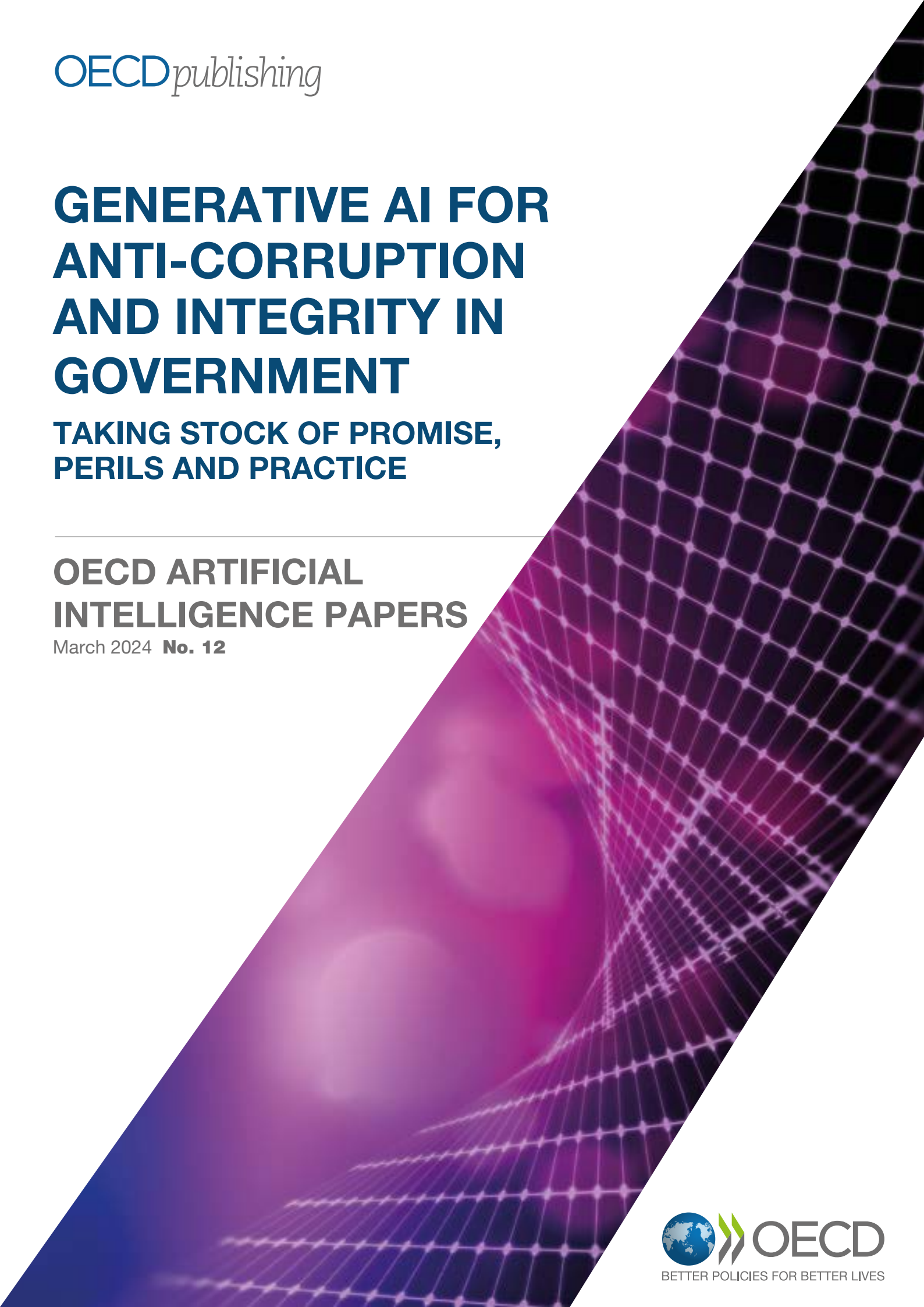
OECD publishing

GENERATIVE AI FOR ANTI-CORRUPTION AND INTEGRITY IN GOVERNMENT

TAKING STOCK OF PROMISE, PERILS AND PRACTICE

OECD ARTIFICIAL INTELLIGENCE PAPERS
March 2024 **No. 12**



OECD
BETTER POLICIES FOR BETTER LIVES

OECD Artificial Intelligence Papers

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors.

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to OECD Directorate for Public Governance, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: gov.contact@oecd.org.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Authorised for publication by Elsa Pilichowski, Director, Public Governance Directorate.

© OECD 2024

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at https://www.oecd.org/termsandconditions/

Foreword

This paper examines the opportunities and challenges related to the use of generative artificial intelligence and large language models reported by a group of government actors engaged in anti-corruption and integrity efforts, termed “integrity actors.” The paper presents insights from responses to a questionnaire administered to integrity actors in government in early 2024.

Under the supervision of Elsa Pilichowski, Director of the Public Governance Directorate (GOV) and the guidance of Julio Bacio Terracino, Head of GOV’s Anti-Corruption and Integrity in Government Division, Gavin Ugale drafted this paper with contributions from Cameron Hall. Jamie Berryhill, Gallia Daor, Claire McEvoy, Mauricio Mejia Galvan, María Pascual Dapena, Karine Perset, Helene Wells and Ricardo Zapata provided valuable comments. Meral Gedik prepared the paper for publication.

The OECD Secretariat would also like to thank the following institutions for providing their insights via the OECD questionnaire: Court of Audit (Austria), Federal Internal Audit (Belgium), Global Affairs (Canada), Office of the Auditor General (Canada), Office of the Comptroller General (Colombia), Office of the Comptroller General (Costa Rica), Ministry of Finance (Costa Rica), Ministry of Justice (Czechia), Agency for Public Finance and Management (Denmark), Office of the Public Prosecutor (Denmark), National Audit Office (Estonia), Ministry of Finance (Finland), Ministry of Justice (Finland), National Bureau of Investigation (Finland), Court of Accounts (France), High Authority for Transparency in Public Life (France), Ministry of Digital Governance (Greece), National Transparency Authority (Greece), Integrity Authority (Hungary), National Tax and Customs Administration (Hungary), Anti-Corruption Authority (Italy), Board of Audit (Japan), Board of Audit and Inspection (Republic of Korea), Corruption Prevention and Combatting Bureau (Latvia), State Audit Office (Latvia), Ministry of Public Administration (Mexico), Ministry of Interior and Kingdom Relations (Netherlands), Court of Audit (Netherlands), Agency for Public and Financial Management (Norway), Government Security and Service Organisation (Norway), Court of Auditors (Portugal), Directorate-General for Administration and Public Employment (Portugal), Court of Audit (Slovenia), the General Comptroller of the State Administration (Spain), National Audit Office (Sweden), Internal Audit of the Embassy of Sweden to Guatemala (Sweden), Federal Statistical Office (Switzerland), National Audit Office (United Kingdom), Department of State (United States), Government Accountability Office (United States), European Court of Auditors (European Union), European Confederation of Institutes of Internal Auditing (European Union), Agency for the Prevention of Corruption and Coordination of the Fight against Corruption (Bosnia and Herzegovina), Office of the Comptroller General (Brazil), Federal Court of Accounts (Brazil), General Inspectorate of the State (Djibouti), Office of the Comptroller General (Ecuador), Integrity and Anti-Corruption Commission (Jordan), Ministry of Economy (Kosovo),[1] National Audit Office (Malta), Agency for Prevention of Corruption (Montenegro), National Anti-Corruption Centre (Republic of Moldova), General-Directorate of Anti-Corruption (Romania), Agency for Prevention of Corruption (Serbia), General Control of Finance (Tunisia), Presidency of the Government (Tunisia), Court of Accounts (Tunisia), State Audit Service (Ukraine).

The OECD Secretariat would also like to express its gratitude to Taka Ariga (Government Accountability Office, United States), Gutemberg Assuncao Vieira (Office of the Comptroller General, Brazil), Máté Benyovszky (Integrity Authority, Hungary) and Emanuele Fossati (European Court of Auditors, European Union) for their insights that helped to shape the questionnaire.

Table of contents

Foreword 3

Abbreviations and acronyms 6

Executive summary 7

1 Generative AI: Opportunities for enhancing anti-corruption and integrity in government 9
1.1. The OECD's questionnaire on generative AI for integrity and anti-corruption 10
1.2. Overview of the maturity of generative AI initiatives 11
1.3. Opportunities and benefits of LLMs for integrity actors 15
Annex 1.A. Key dimensions for assessing institutional digital maturity 21

2 Generative AI: Challenges, risks and other considerations for integrity actors in government 25
2.1. Overview of main challenges for integrity actors to adopt generative AI and LLMs 26
2.2. Building a generative AI and LLM capacity within institutions responsible for integrity and anti-corruption 31
2.3. Ensuring the responsible development and use of generative AI and LLMs by integrity actors 34
2.4. Mitigating the risk of generative AI as a tool to undermine integrity 43

References 46

Notes 49

FIGURES

Figure 1.1. Number of respondents to the OECD's questionnaire by type of organisation 10
Figure 1.2. Stage of generative AI and LLM use by type of organisation 12
Figure 1.3. Maturity of generative AI and LLM use by region 13
Figure 1.4. Perceived benefits of generative AI and LLMs for integrity actors' internal operations 17
Figure 1.5. Perceived benefits of generative AI and LLMs for anti-corruption activities by type of organisation (top two choices) 20
Figure 2.1. Main challenges for deploying generative AI and LLMs 26
Figure 2.2. Main challenges for deploying generative AI and LLMs by type of organisation 27
Figure 2.3. Primary data sources for building LLMs among questionnaire respondents 30
Figure 2.4. Integrity actors' approach for using LLMs 31
Figure 2.5. Safeguards to ensure responsible use of AI and LLMs 35
Figure 2.6. GAO's Artificial Intelligence Accountability Framework 40

BOXES

Box 1.1. The government-wide vision on generative AI of the Netherlands 14
Box 1.2. Lessons from Brazil's SAI and the development of ChatTCU 18
Box 2.1. The generative AI training programmes of the European Court of Auditors (ECA) 28
Box 2.2. Retrieval-Augmented Generation for LLMs 32
Box 2.3. France's LLaMandement for summarising legislative text 33
Box 2.4. The Office of the Comptroller General (CGU) of Brazil's approach to piloting LLMs 34
Box 2.5. The Corruption Prevention Commission (CPC) of Armenia's use of AI to verify asset declarations 37
Box 2.6. The OECD Principles on Artificial Intelligence 38
Box 2.7. The AI Accountability Framework of the US Government Accountability Office (GAO) 40
Box 2.8. Human-centred considerations for promoting transparency when evaluating LLMs 42
Box 2.9. Insights from the Independent Commission Against Corruption (ICAC) of New South Wales on AI's potential threats to anti-corruption work 43

Abbreviations and acronyms

ACA	Anti-Corruption Agency
AI	Artificial Intelligence
GPT	Generative Pre-trained Transformers
IT	Information Technology
LLM	Large Language Model
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
RAG	Retrieval-Augmented Generation
SAI	Supreme Audit Institution

Executive summary

Generative artificial intelligence (AI) has been part of the technological landscape for some time, but recent developments, particularly in large language models (LLMs) as one type of generative AI, have recently propelled it into a position of disruptive influence. Governments must keep pace with this innovation not only as regulators, but also as users. This paper explores the latter challenge with a focus on integrity actors, including anti-corruption agencies (ACAs) and oversight bodies, such as supreme audit institutions (SAIs) and internal audit functions.

The integrity actors who offered insights for this paper identified several opportunities and benefits of generative AI, focusing largely on their exploration and use of LLMs. For instance, integrity actors in Brazil are deploying LLMs to sift through massive datasets to identify patterns indicative of fraud, offering insights for investigations and risk mitigation measures. Integrity actors in Finland, France, Greece and the United Kingdom are using LLMs to support in drafting documents, analysing spreadsheets and summarising texts. These LLMs can make the day-to-day work of auditors and investigators more efficient, thereby freeing them from time-consuming organisational tasks.

Integrity actors also highlighted various challenges, ranging from technical ones concerning the integration of LLMs to strategic questions about ensuring trustworthy AI systems. Integrity actors recognise that LLMs are an evolving technology capable of "hallucinations," whereby they may generate convincing yet inaccurate, fabricated or misleading information, based on unclear reasoning. This inherent complexity in how LLMs generate outputs can perpetuate a lack of transparency and accountability in decision making, which can undermine the very principles that integrity actors seek to uphold. Failure to mitigate these risks, curb bias and promote responsible and ethical use of AI, can have harmful real-world impacts, such as the reinforcing of structural inequalities and discrimination.

To identify and explore these opportunities and challenges, the OECD sent a questionnaire to and interviewed organisations from several OECD communities, including the Working Party of Senior Public Integrity Officials, the Auditors Alliance, and a Community of Practice on Technology and Analytics for Public Integrity. Based on the responses of 59 organisations from 39 countries, the OECD collected key insights concerning the use of generative AI and LLMs. They included the following:

- Generative AI, particularly LLMs used for processing and generating text, can enhance the internal operations of integrity actors, with the most promising gains in operational efficiency and analysing unstructured data. For investigative and audit processes, integrity actors saw the highest value of LLMs in evidence gathering and document review, with a significant portion of respondents, especially those conducting performing audits, prioritising these activities.
- LLMs show promise for strengthening several anti-corruption and anti-fraud activities, but examples in government are limited and the return on investment is unclear. Integrity actors viewed document analysis and text-based pattern recognition as the most valuable use cases of LLMs for anti-corruption and anti-fraud. However, respondents reported few advanced initiatives in this area, and many organisations are still incubating ideas.

- Integrity actors cited a shortage of skills and IT limitations as the biggest challenges they face to implement LLMs. Many institutions expressed that they either lack sufficient financial, human, and technical resources to deploy LLMs entirely, or their staff does not have sufficient data literacy to use such tools. Concerns about budget constraints were comparatively more pronounced among internal audit bodies and ACAs relative to SAIs.
- Advice for piloting LLMs includes first incorporating them into low-risk processes and considering the requirements for scaling early on. Such an approach can build capacity where mistakes are not as costly before scaling generative AI to riskier, more resource-intensive and more analytical tasks. Having an early handle on the organisational needs for computational and storage resources can help an organisation to prepare for scaling.
- Integrity actors mostly rely on turnkey foundation LLMs developed by technology companies. Various options exist to develop LLMs, from open-source models to those created by private firms or government entities. In practice, integrity actors that responded to the questionnaire are either using an existing, turnkey model outright or they are fine-tuning a foundation model (i.e. further training a pre-trained LLM with specific datasets to adapt its capabilities for particular tasks).
- Overcoming language barriers inherent in using or fine-tuning off-the-shelf LLMs is a key challenge. Currently, most LLMs are trained in English, which poses limitations for many integrity actors who wish to deploy models in their native language. To address this challenge, some countries are investing in the development of local language LLMs.
- Integrity actors recognised the need for safeguards in some areas but can do more to ensure trustworthy AI systems, as well as the responsible and ethical use of generative AI as initiatives mature. Integrity actors can improve their focus and activities to mitigate the risks of bias and discrimination and address ethical concerns in how they use and apply LLMs internally.
- Integrity actors can put a greater emphasis on monitoring and evaluating LLMs, including considerations pertaining to the interpretability and explainability of a model's outputs. Evaluating LLMs and attempting to explain results poses complex challenges. However, addressing these challenges with multi-faceted solutions will be critical for the uptake of LLMs amongst integrity actors.
- Generative AI can enhance the work of integrity actors, but it also necessitates greater vigilance of evolving integrity risks. For instance, LLMs provide new ways for integrity actors to operate and assess risks, but they also can accelerate and amplify certain types of fraud and corruption.

The findings from the OECD's questionnaire are not generalisable to all integrity actors. Nonetheless, the paper describes common challenges and potential use cases that are transferable across contexts, providing inspiration to integrity actors as they consider how to make the most of this rapidly evolving technology. The OECD's policy-focused work offers inspiration throughout the paper, including the work of the OECD.AI Policy Observatory, as well as the OECD's Recommendation of the Council on Artificial Intelligence and the Recommendation on Digital Government Strategies.

1 Generative AI: Opportunities for enhancing anti-corruption and integrity in government

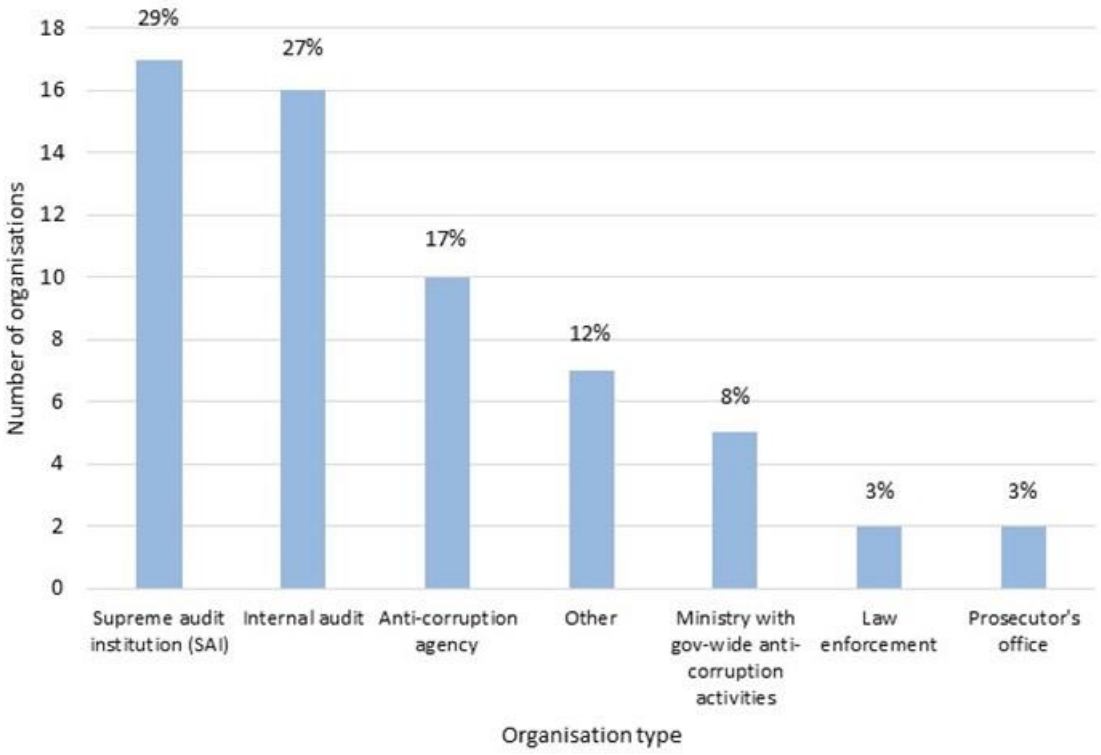
This section explores the opportunities for integrity actors to use generative AI, particularly LLMs, to enhance their internal operations as well as their anti-corruption activities. It presents the views of 59 integrity actors captured in an OECD questionnaire on generative AI for integrity and anti-corruption, including insights into the potential benefits the technology offers. The supreme audit institutions that responded to the questionnaire are generally the most advanced in their use of generative AI among questionnaire respondents. However, most integrity actors that responded to the questionnaire are still in the early stages of thinking about or developing generative AI tools.

1.1. The OECD's questionnaire on generative AI for integrity and anti-corruption

In January 2024, the OECD administered a questionnaire for integrity actors in government on the use of generative AI for public integrity and anti-corruption. To implement the questionnaire, the OECD relied primarily on three of its communities: the Working Party of Senior Public Integrity Officials, the Community of Practice on Technology and Analytics for Integrity and the Auditors Alliance. With the help of members of these communities, the OECD identified integrity actors in government with the relevant mandate and expertise to react to a questionnaire about generative AI for integrity and anti-corruption. Several participants from the Community of Practice piloted the questionnaire and select respondents provided additional insights via targeted interviews.

For purposes of the questionnaire and this paper, integrity actors include anti-corruption agencies (ACAs), supreme audit institutions (SAIs), internal audit or control functions, and ministries with government-wide integrity and anti-corruption activities (e.g. Ministry of Public Administration). They also include law enforcement and prosecutors' offices. These integrity actors together account for 88% (52) of the 59 organisations that responded to the questionnaire.[2] All but one of the other seven organisations to respond represent government entities that are responsible for AI policy. One Tax and Customs Administration responded to the questionnaire as well. SAIs and internal audit functions provided just over half of all responses. Figure 1.1 summarises some of these key features of the organisations that responded.

Figure 1.1. Number of respondents to the OECD's questionnaire by type of organisation



Note: The percentages show the proportion of organisations out of a total 59 that responded to the questionnaire. Internal audit bodies include both central internal audit bodies and internal audit units of individual institutions, as well as comptroller general's offices and ministries and agencies responsible for public financial management. The "other" category contains primarily ministries responsible for government-wide AI policy as well as one tax and customs agency.
Source: OECD questionnaire

Respondents to the questionnaire represent a broad range of government entities with different institutional mandates with regards to public integrity and anti-corruption. Most of the respondents have roles and responsibilities related to IT, data science, AI or digital initiatives within their organisation. The OECD did not attempt to identify or contact the entire sub-populations of integrity actors, as we define them in this paper. The ultimate purpose of this paper and the questionnaire is to explore current use cases and provide a snapshot of practices, opportunities and challenges. As such, the results are not generalisable to broader populations. All descriptive statistics that illustrate key findings reflect responses to the OECD's questionnaire without exception.

1.2. Overview of the maturity of generative AI initiatives

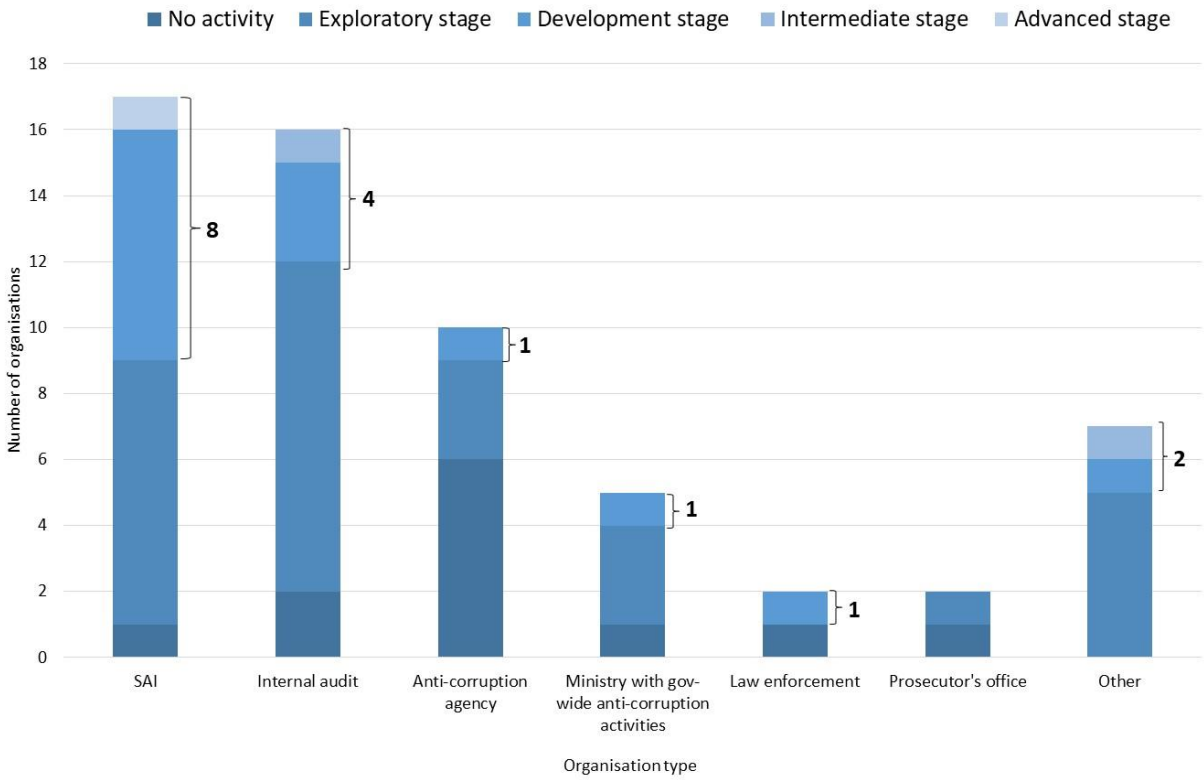
In recent years, generative AI surged in prominence with the rise of deepfakes and the introduction of transformative models like Generative Pre-trained Transformers (GPTs) and other large language models (LLMs), marking a significant leap forward in the field. LLMs are advanced machine learning algorithms proficient in interpreting inquiries or commands and producing responses in human-like language. These models function by processing extensive datasets during their training phase, allowing them to identify statistical correlations, such as how words relate to each other and the contextual importance of words within sentences. Utilising this insight, the models are capable of sequentially generating text, predicting each subsequent word in a sequence (OECD, 2023[1]) (Shabsigh and Boukherouaa, 2023[2]). The technology captured global interest in November 2022 with the introduction of text-to-image generators and the release of Open AI's ChatGPT (Lorenz, Perset and Berryhill, 2023[3]).

In this context, integrity actors have had little time to comprehend the opportunity generative AI presents for their work, let alone to fully integrate it into activities. When the OECD surveyed integrity actors, the expectation was that across the board the respondents would describe their organisations as being in the early stages of maturity concerning the use of generative AI and LLMs. Not only is the technology relatively new, but government entities—integrity actors included—are not known for being first movers in terms of technology adoption. The responses to the OECD's questionnaire reflect these expectations. Of the 59 organisations that responded from 39 countries, as well as two supranational organisations in the European Union, approximately 50% (30) reported they do not use generative AI in their operations, but they are exploring potential use cases. Another 24% (14) of respondents indicated their institutions are in the development phase. In other words, they have experimented with generative AI in a few projects, but it is not yet integrated into the organisations' operations.

SAIs' efforts to use generative AI were the most mature relative to other types of organisations, including one respondent who described their SAI's use of the technology as "advanced." Overall, 47% (8) of SAI respondents reported being at least in the development stage of using generative AI, the highest percentage of the different organisational types. After SAIs, 25% (4) of respondents working in internal audit bodies said their organisation is in the development stage or beyond, while only one institution in each of the other categories has reached at least the development stage. Figure 1.2 summarises these results and provides definitions for the different stages. The counts highlighted with brackets indicate the number and type of organisations that have reached at least the development stage, which is a subgroup of surveyed organisations that is the focus of subsequent analysis.

Figure 1.2. Stage of generative AI and LLM use by type of organisation

Which of the following options best describes the maturity of your institution's use of Gen AI and LLMs specifically, as a sub-domain of AI?



Note: The data label callouts highlight the number of institutions that have reached at least the development stage. Internal audit bodies include both central internal audit bodies and internal audit units of individual institutions, as well as comptroller general's offices and ministries and agencies responsible for public financial management. The "other" category contains primarily ministries responsible for government-wide AI policy as well as one tax and customs agency. Possible responses included the following: 1) Advanced Stage: Gen AI is deeply integrated into our core operations and we continuously seek ways to improve and expand its use. 2) Intermediate Stage: Gen AI is used in several areas of our activities, but it is not yet fully optimised or widespread. 3) Development Stage: We have experimented with Gen AI in a few projects but it is not yet integrated into our operations. 4) Exploratory Stage: We do not use Gen AI in our operations but we are currently exploring potential uses. 5) No Activity: We do not use Gen AI in our operations and we are currently not exploring potential uses.
Source: OECD questionnaire.

While these results are not generalisable, they align with the OECD's experiences working with these communities. Among SAIs that have successfully incorporated innovative approaches to the use of technology, data, and analytics into their audit work, a common thread is their openness to experimentation. In some countries, SAIs may also have access to more resources than other types of integrity actors, therefore enabling more experimentation, as highlighted later in the paper. This commitment to experimentation remains consistent even when other aspects of the SAI's work and culture tend to be risk averse. For those SAIs that have established dedicated "Innovation Labs," experimentation has become a strategic objective.

One notable advantage of an innovation lab is its role in institutionalising knowledge and expertise. This model can help to advance new methodologies that can benefit multiple departments within the SAI. For example, SAIs in countries like Brazil, the United States, and Norway have all established effective innovation labs to assist auditors in keeping pace with technological developments and drive continuous professional development (OECD, 2022[4]). This includes the integration of technology and data-driven

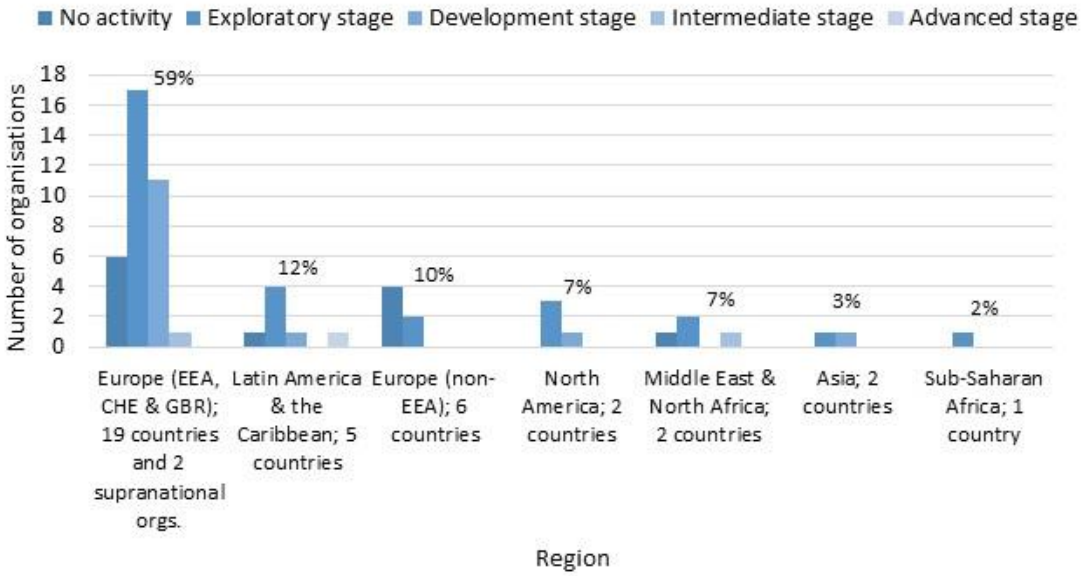
approaches into their auditing processes, as well as enhancing their knowledge for auditing emerging areas in government, such as the deployment of AI.

In all, 59 organisations from 39 countries responded to the OECD's questionnaire. Respondents predominantly represented European countries, including 19 countries from the European Economic Area (EEA), Switzerland and the United Kingdom (UK), 2 supranational organisations in the European Union, as well as six non-EEA countries (i.e. from EU candidate and neighbourhood countries). Of the organisations representing the countries from the EEA, including the two supranational organisations, 34% (12) reported having reached the development stage (11) or intermediate state (1), although this number was 0% in non-EEA, European countries (see Figure 1.3). Among the five countries represented in the responses from Latin America, the two organisations that indicated a level of maturity at the development and advanced stages were both from Brazil. The only organisation of the 59 respondents that reported an advanced stage of generative AI and LLM use was based in Brazil. In other regions, while the number of responses was low, countries generally were in the exploratory or development stages.

This figure is not meant to allow for drawing comparisons about the digital maturity of integrity actors in different regions. As noted, the questionnaire only covered a subset of the global population of integrity actors, so any conclusions about digital maturity are only representative of the pool of respondents. Judging from exchanges between the OECD and integrity actors during the course of this analysis, it is likely that many other organisations that received the questionnaire chose not to respond because they did not have any activities or discussions concerning the use of generative AI whatsoever.

Figure 1.3. Maturity of generative AI and LLM use by region

Which of the following options best describes the maturity of your institution's use of Gen AI and LLMs specifically, as a sub-domain of AI? (*The percentages refer to the proportion of organisations from each region out of a total 59 organisations that responded to the questionnaire.*)



Note: "EEA" is European Economic Area, "CHE" is the country code for Switzerland and "GBR" is the country code for the United Kingdom. Possible responses included the following: 1) Advanced Stage: Gen AI is deeply integrated into our core operations and we continuously seek ways to improve and expand its use. 2) Intermediate Stage: Gen AI is used in several areas of our activities, but it is not yet fully optimised or widespread. 3) Development Stage: We have experimented with Gen AI in a few projects but it is not yet integrated into our operations. 4) Exploratory Stage: We do not use Gen AI in our operations but we are currently exploring potential uses. 5) No Activity: We do not use Gen AI in our operations and we are currently not exploring potential uses.
Source: OECD questionnaire.

Responses to the questionnaire suggest that digital maturity is higher concerning the broader use of AI than it is for generative AI specifically, suggesting that countries are employing strategic approaches to exploring and deploying AI use. Specifically, around 34% (20) of organisations are currently developing a strategy for the use of AI in their institution, while several others follow a government-wide strategy for the use of AI. Six institutions currently have an AI strategy in place, all of which were from EU countries with one exception.

The efforts of these integrity actors illustrate the value they place on formally recognising the need for a strategic approach to exploring and deploying AI. Having a digital strategy with clear goals, objectives, performance indicators and defined roles and responsibilities, among other features, is a critical aspect of digital maturity (see Annex 1.A), and an AI strategy is often a subset of such a digital strategy. As one example, Box 1.1 describes the efforts of the Netherlands to incorporate generative AI into its broader AI strategy as well as the work of public bodies, including those responsible for anti-corruption. Generative AI can also be incorporated into the strategies of specific institutions. For example, Norway's Office of the Auditor General (OAG) envisions increased use of AI in performance audits in its 2018-2024 Strategic Plan (Office of the Auditor General of Norway, 2018[5]).

Box 1.1. The government-wide vision on generative AI of the Netherlands

The Netherlands became one of the first countries to publish a strategy focused specifically on generative AI in January 2024. The government-wide vision on generative AI outlines the opportunities and challenges posed by generative AI, elaborates a vision for the use of generative AI in the public sector based on four principles, and establishes specific actions to ensure public sector generative AI use is responsible and effective. This strategy provides an example of how integrity actors can benefit from a broader strategic approach to generative AI in the public sector.

The four principles to guide the development of generative AI, as outlined in the strategy, are as follows:

1. Generative AI is developed and applied in a safe way
2. Generative AI is developed and applied equitably
3. Generative AI that serves human welfare and safeguards human autonomy
4. Generative AI contributes to sustainability and prosperity

Opportunities discussed in the strategy include generative AI's potential to automate administrative and legal processes, serve as a learning tool, and even solve problems requiring complex data analysis with many inputs. On the other hand, risks include the impact on citizens relating to bias and privacy, increased dependence on foreign tech companies with monopoly power, exacerbating job insecurity, and the proliferation of mis- and disinformation. Both sides of this issue are relevant for integrity actors. For example, the amount of complex data analysis required of many of these actors means that generative AI presents notable opportunities, while the sensitivity of this data means that risk mitigation is also necessary.

Moreover, some of the actions laid out in the strategy explicitly highlight the participation of integrity actors. For instance, the strategy advocates for pre-deployment audits of advanced models and assigns an action to the Ministry of Foreign Affairs to promote this practice—along with responsible use of generative AI more broadly—on the international stage. The action plan envisions using generative AI for legal and administrative processes and analysing large datasets, which would be relevant for integrity actors. Beyond this, since the Netherlands is taking a whole of government approach, all actions taken will support the responsible deployment of generative AI in integrity bodies as a subset of the public administration.

Source: (Government of the Netherlands, 2024[6])

1.3. Opportunities and benefits of LLMs for integrity actors

The OECD supports integrity actors in government to build their technological capacity and develop data-driven methodologies for assessing fraud and corruption risks. The digital maturity of these partner organisations varies widely, with a small group implementing advanced analytics and a larger group relying more on qualitative risk assessments. The work of other organisations reflects a similar reality where risk assessments typically involve manual analysis, which can be time-consuming, resource-intensive and inefficient, often relying on specific complaints or anecdotes (World Bank, 2023[7]). Advancements in the ability of governments to harness technology, data and analytics, as well as ever-evolving AI methodologies, are challenging this status quo.

While it may not be the norm, integrity actors in the public sector have for years successfully leveraged advanced analytics and AI, such as supervised machine learning, to uncover hidden patterns and anomalies that indicate potential corrupt or fraudulent behaviour. For instance, supervised machine learning helped the General Comptroller of the State Administration of Spain (*Intervención General de la Administración del Estado*, IGAE) to detect fraud and corruption by leveraging proven cases as training data, enabling the model to learn and identify complex patterns and anomalies in public grants indicative of fraud (OECD, 2021[8]). OECD members and partners across the globe, including public integrity partners from Brazil, Colombia, Korea, Lithuania and the United States, are advancing similar efforts (OECD, 2022[4]; OECD, 2021[8]). AI and data-driven assessments enable organisations to proactively mitigate risks and safeguard taxpayer money in ways that are more efficient and impactful than more manual approaches, while allowing for wider covering of the risk universe.

1.3.1. There are a variety of applications for LLMs in the integrity and anti-corruption space

The advent of generative AI, and in particular LLMs, creates new avenues for integrity actors to enhance the efficiency and impact of their work. This paper provides examples of some of these opportunities, which broadly cover two dimensions: the organisation's internal operations, and more specifically, anti-corruption and anti-fraud activities. Based on responses to the OECD's questionnaire, LLMs are a main focus of integrity actors' current exploration with generative AI, so much of the paper concentrates on this technique.

LLMs are well-suited to support integrity actors in automating certain fraud detection activities, such as querying documents and data sources for potential risk. LLMs can also help auditors and investigators to carry out many operational tasks that, while not unique to integrity actors, are particularly promising given the high volumes of documentation and data that audit, anti-corruption and investigative bodies typically process. For instance, LLMs can help to organise large volumes of text for easier prioritisation and consumption, and aid in root-cause analyses or pattern recognition (U.S. Government Accountability Office, 2024[9]). Some countries such as Sweden are developing government-wide virtual assistants that would help streamline these operational tasks in all public bodies, including integrity bodies (AI Sweden, 2024[10]). The efficiencies gained by these techniques can reduce both effort and error, allowing auditors and investigators to focus more on high-value tasks that require human judgement and expertise, which generative AI has yet to replace. By making anti-corruption and anti-fraud activities more effective, generative AI can also strengthen public integrity and accountability.

Academia offers additional inspiration for integrity actors to apply LLMs. For instance, financial and accounting literature provides numerous examples of using LLMs to assess financial texts. One group developed an LLM called FinBERT, based on Google's Bidirectional Encoder Representations from Transformers (BERT) algorithm and a large corpus of financial texts, for sentiment analysis and extracting specific discussions about environment, social and governance (ESG) (Huang and Yi Yang, 2023[11]). Another group of researchers took a case study approach and explored the adoption of ChatGPT by a

multinational company's internal audit function (IAF) across various stages of the audit process, including risk-based audit planning, audit preparation and data analysis. In this instance, the IAF observed promising results in tasks that involved scoping audits, brainstorming risks, drafting descriptions, interview preparation and report writing (Emett, 2023[12]). These texts highlight opportunities, but they also warn of risks and elaborate on challenges of deploying LLMs, some of which are covered in Section 2.

LLMs also have the potential to promote integrity in public spending if adopted by a broader range of actors that do not fit the definition of integrity actors used for this paper. For instance, LLMs, such as those that power ChatGPT, can support public procurement officials in analysing large amounts of data on a company and potential contractor to screen for fraud or corruption risks. One organisation that responded to the questionnaire highlighted the development of a pilot project to continuously identify risk indicators in public procurement processes using LangChain and an LLM to preprocess the unstructured data.[3] The organisation executes the preprocessing phase centrally, while leveraging the expertise of auditors in a more decentralised manner to provide prompts that pinpoint procurement features of interest.

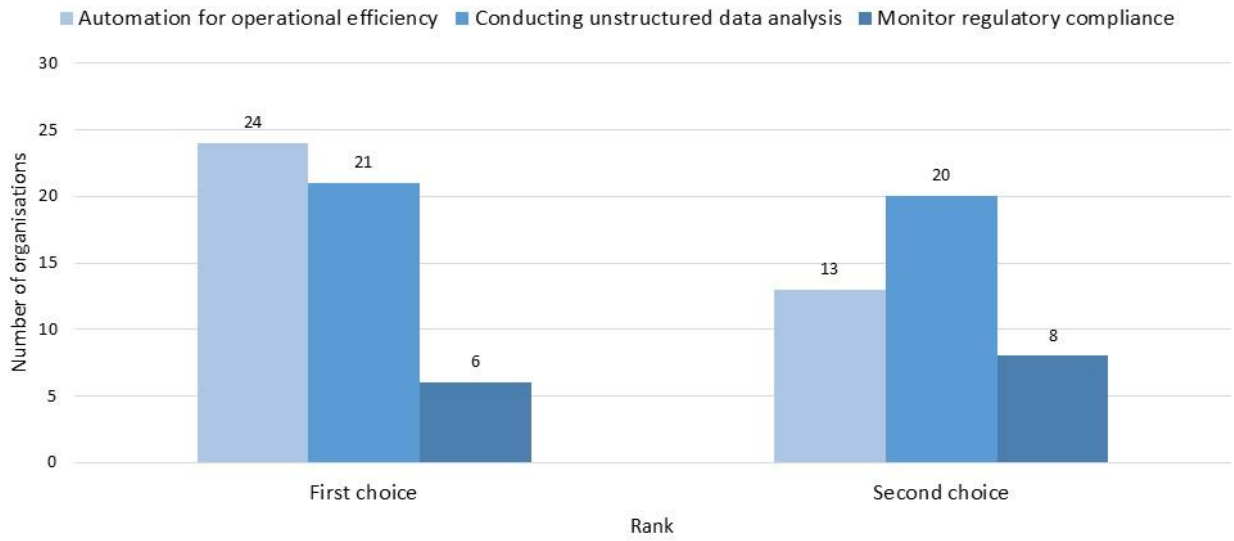
1.3.2. LLMs can enhance the internal operations of integrity actors, with gains in operational efficiency and analysing unstructured data being the most promising opportunities

The OECD asked integrity actors where they think generative AI, LLMs in particular, can add the most value for their organisation and its activities. As noted, the question focused on opportunities in two areas: 1) the organisation's internal operations; and 2) its anti-corruption and anti-fraud activities. These areas, as well as the options for responses, are difficult to artificially separate and they may not be mutually exclusive. Gains in one area of analysis or information processing can lead to efficiencies in others. With that in mind, OECD questionnaire respondents ranked operational efficiency and unstructured data analysis as the areas that could benefit the most from the use of generative AI and LLMs for internal operations (the first choice of 45 of the 59 organisations, or 76%). Figure 1.4 shows the areas of added value ranked first and second, with additional information about the various areas of internal operations surveyed.

A small group of integrity actors ranked monitoring of regulatory compliance as the main area of perceived value of generative AI and LLMs. Respondents viewed the contributions of generative AI to public engagement and transparency and training and capacity building as comparatively smaller. When breaking down the data by organisational type, over half of the 16 SAIs (9) that responded to the questionnaire ranked unstructured data analysis as the number one potential benefit of LLMs, but there were no other significant trends or patterns in the data by organisational type.

Figure 1.4. Perceived benefits of generative AI and LLMs for integrity actors' internal operations

Within your institution, which of the following areas of internal operations would benefit the most from the use of Gen AI and LLMs?



Notes: Possible responses included the following: 1) Operational Efficiency: Streamlining internal processes by automating routine tasks for core activities, allowing for more efficient allocation of human resources. 2) Unstructured Data Analysis: Leveraging Gen AI to effectively analyse and interpret unstructured data, such as text, images, and audio, which can provide deeper insights and inform decision-making processes. 3) Public Engagement and Transparency: Using LLMs to streamline communication with the public and stakeholders. 4) Training and Capacity Building: Using LLMs for training purposes, such as planning curricula and workshops. 5) Regulatory Compliance Monitoring: Employing Gen AI to continuously monitor and ensure compliance with relevant laws and regulations, reducing the likelihood of non-compliance issues. 6) Not sure. 7) Other.
Source: OECD questionnaire

Respondents also offered their views about the value of generative AI, includingLLMs, for investigative and audit processes. They ranked gathering evidence and document review as having the highest value in this respect. Specifically, 37% of respondents (22) ranked these activities as their top choice, followed by the use of generative AI and LLMs for selecting audits and investigations (ranked first by 25%, or 15 organisations). This was particular the case among SAIs and internal audit bodies, which as a group, ranked evidence gathering and document review higher relative to other integrity actors. As far as the value of generative AI and LLMs for other audit and investigative activities, fewer integrity actors ranked the following options at the top: drafting reports and producing graphics (7); none of the above/not sure (6); planning audits and investigations (5); generating content for public relations (3); and documenting processes (1). The initiative of Brazil's SAI (*the Tribunal de Contas da União*, TCU) to develop ChatTCU illustrates one approach for leveraging LLMs to enhance the efficiency with which auditors gather and review documentation. Box 1.2 describes the initiative and offers key lessons learned, many of which are broadly applicable to other types of organisations, even though ChatTCU is an SAI-led initiative.

Box 1.2. Lessons from Brazil's SAI and the development of ChatTCU

In February 2023, the Brazilian Federal Court of Accounts (TCU) launched ChatTCU based on OpenAI's ChatGPT. The TCU built the tool based on the view that LLMs are not a passing trend, and therefore it decided to take an institutional approach to consciously developing use cases while addressing potential risks. While the initiative is still developing, the TCU has already experimented with several applications. As of December 2023, the TCU reported over 1 400 users, demonstrating the extent to which the tool has been rolled out and adopted.

The current version of ChatTCU is integrated with TCU's systems, providing answers based on the Court's cases, selected precedents, and administrative system, coupled with the knowledge base of ChatGPT itself. For instance, ChatTCU allows auditors to request a summary of a case document, pose technical questions related to TCU's work and court decisions, and seek help for administrative services. ChatTCU v3 is based on GPT-4 32k, which grants better quality to the answers provided and fewer chances of errors or hallucinations.

The TCU plans to incorporate a range of new features, such as further integration with other systems and workflow automation through user prompts. TCU hosts ChatTCU on a dedicated instance of Microsoft Azure's cloud platform. This helps TCU to ensure the security and confidentiality of its data, and it allows auditors to use the tool without sending private data to OpenAI. Furthermore, hosting ChatTCU in this way helps facilitate integration with other systems. Key lessons learned from the TCU's experience, many of which are transferable to other integrity actors include:

- **Internalise technology**: The proactive development of ChatTCU, tailored to TCU's needs, suggests that integrity actors could consider building their own AI solutions rather than relying solely on external tools. This also helps build the digital literacy and capacity within the organisation that will prove valuable in other areas.
- **Integration with existing systems:** The integration of ChatTCU with TCU's existing systems allowed auditors to access administrative information and gain insights into audits more efficiently, underscoring the importance of seamless integration with existing workflows and systems.
- **Scalability and future-proofing:** TCU's plans to expand ChatTCU's functionalities demonstrate the need for scalability and adaptability in AI solutions, urging integrity actors to plan for future upgrades and developments.
- **Potential for standardisation:** TCU's consideration of incorporating audit standards into ChatTCU indicates the potential for AI tools to assist in maintaining standards, suggesting that other integrity actors may explore similar possibilities to enhance their processes.
- **Feedback-driven development**: TCU underscored the importance of collecting user feedback to continuously improve AI solutions, emphasising the need for integrity actors to create mechanisms for staff to provide feedback and suggestions for enhancements.
- **Multidisciplinary approach:** TCU formed a multidisciplinary working group to assess the risks and opportunities of using generative AI, involving representatives from various areas and promoting debates to help make informed decisions about AI implementation.
- **Invest in training and awareness**: TCU's emphasis on raising awareness among staff about the potential and risks of AI highlights the crucial need for training and educating staff members on how to effectively use AI technologies. Involving staff in developing AI solutions internally will also help them learn how to tackle these challenges firsthand.

Source: Responses to the OECD's questionnaire and https://portal.tcu.gov.br/en_us/imprensa/news/chattcu-integration-of-the-tool-into-the-courts-systems-improves-the-use-of-generative-artificial-intelligence-in-external-control-activities.htm

1.3.3. Generative AI and LLMs show promise for strengthening a variety of anti-corruption and anti-fraud activities, but examples in government are limited and the return on investment is unclear

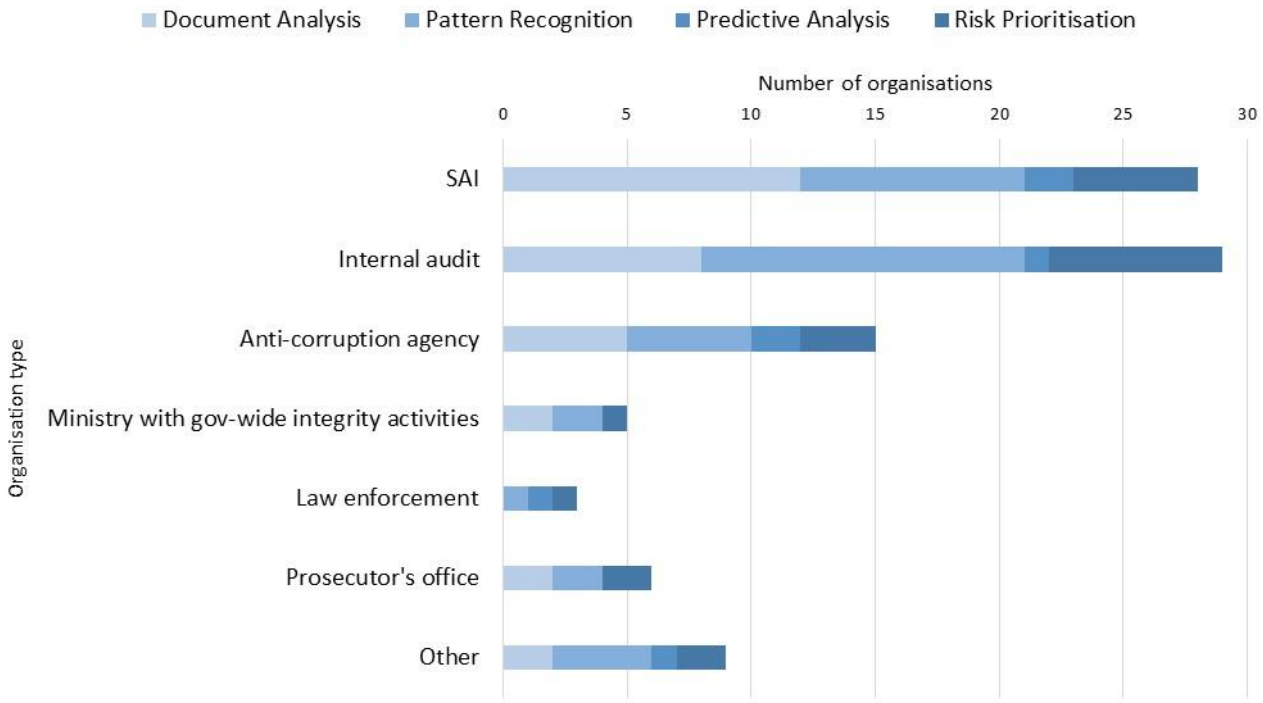
The OECD also asked respondents for their views on the perceived value of generative AI, including LLMs, for anti-corruption and anti-fraud activities. Document analysis and pattern recognition had the highest perceived value with most respondents ranking these activities as either their first or second option. They ranked risk prioritisation as third (15%, or 9 respondents). Following those top 3 selections, in smaller numbers, respondents chose either none of the above/not sure, developing training and simulation tools and conducting predictive analytics to prevent and detect fraud and corruption. ACAs showed the most interest in pattern recognition, with half of the 10 ACAs that responded to the questionnaire ranking this at the top of the list of perceived value of generative AI and LLMs. Several questionnaire respondents highlighted the fact that their mandate does not allow them to conduct anti-fraud activities.

Figure 1.5 summarises the top four responses by type of organisation for the perceived benefits of generative AI for anti-corruption and anti-fraud activities, and it provides definitions for different activities in the questionnaire. Some of the activities could overlap in practice and raise additional questions for future inquiry and research. For instance, activities to detect patterns and anomalies, which many respondents ranked as their top choice, could also inform risk prioritisation, which generally was ranked lower by most respondents. There could be practical reasons for this. Experimenting with LLMs and analysing unstructured data draws from finite resources. OECD's projects with integrity actors demonstrate that many organisations are already investing resources in response to other technological trends (e.g. "big data" analysis), including how to make better use of structured data for assessing risks.

Moreover, as an activity, risk prioritisation is an obvious candidate for experimenting with generative AI, but as an institutional process, it can already have its own set of established procedures, processes and tools in place. These are developed and refined over decades. In this context, whether it is an SAI, ACA, law enforcement body or other integrity actor, any new LLM-supported methodology would need to be thoughtfully designed and integrated if it is to become part of, or potentially disrupt, the status quo. This poses challenges that are not just technical in nature, but also organisational, legal and even political (internally), which may help to explain why many respondents did not rank risk prioritisation higher.

Figure 1.5. Perceived benefits of generative AI and LLMs for anti-corruption activities by type of organisation (top two choices)

Within your institution, which of the following anti-corruption or anti-fraud activities would benefit most from the use of Gen AI and LLMs?



Note: The numbers in the chart refer to the number of times an institution mentioned each activity as either their first or second ranked choice. Possible responses included the following: 1) Pattern Recognition: Identifying unusual patterns or anomalies in data that may indicate corrupt or fraudulent activities. 2) Document Analysis: Automating the review of large volumes of documents for potential corruption or fraud indicators. 3) Risk Prioritisation: Assisting in risk assessment and prioritisation of investigations based on AI-generated insights. 4) Predictive Analysis: Using LLMs for predictive analytics to anticipate and prevent potential corrupt or fraudulent activities. Other options not illustrated include: 5) Training and Simulation: Providing training and simulation tools to staff for better understanding and detection of corruption/fraud; 6) None of the above/not sure; and 7) Other.
Source: OECD questionnaire

One respondent with government-wide integrity and anti-corruption activities (i.e. a Ministry of Justice) highlighted the processing and review of asset declarations as one specific area of need and potential value of LLMs. Echoing the OECD's experience with many organisations responsible for asset declaration systems, the respondent described a high-volume of checks required for verifying the content of asset declarations. These checks are currently done manually in most countries. Having a tool to enhance the processing of these declarations is not just about creating more efficient processes and procedures. Such solutions would ultimately contribute to greater transparency in government and enhance public awareness about conflicts-of-interest concerning public officials.

Annex 1.A. Key dimensions for assessing institutional digital maturity

The dimensions and key practices below are based on reviews of academic literature, discussions with subject matter experts in government, industry and non-governmental institutions, insights from the OECD's technical support for governments to strengthen their digital strategies and data-driven risk assessments, as well as OECD Recommendations.[4] Numerous self-assessment tools for digital or technology readiness that are relevant or made for integrity actors at an institutional level, such as the Supreme Audit Institutions (SAI) Information Technology Maturity Assessment, also provided inspiration. These practices consider digital maturity and transformation from an organisational perspective, but many are applicable for designing and implementing digital projects.

Strategy and organisation

This dimension encompasses leadership's vision and strategy for digital transformation, including its goals for strengthening the use of digital technologies and data. A digital transformation strategy can stand alone or be integrated with existing organisational strategies. Either way, the aim is to ensure alignment of the digital strategy with other organisational priorities, audit processes and IT strategies (Bumann and Peter, 2019[13]). Clear delineation of roles and charting out responsibilities is imperative. This can include establishing an entity internally with an organisation-wide mandate to implement and co-ordinate digital initiatives. Data management and data governance are also key aspects of this dimension. They involve the policies, procedures, standards and controls that ensure data privacy, quality, consistency and security. Relative to project-based improvements, digital transformation by nature has a disruptive effect on an organisation's traditional approaches to data management and data governance.

Key practices in this dimension include:

- Align the Digital Strategy with the goals and objectives of other institutional strategies, such as the Strategic Plan and IT Strategy.
- Conduct assessments and establish a baseline for digital maturity, capabilities, IT infrastructure and architecture, and possible gaps.
- Identify key opportunities and challenges concerning data management and data governance, including priorities for ensuring data security and quality.
- Define roles and responsibilities internally, including the designation of an entity to implement the Digital Strategy that has an organisation-wide mandate and access to leadership.
- Engage with key stakeholders in the design of the Digital Strategy, including leadership, management and users of new tools, to understand digital maturity and priorities.
- Establish a plan and key performance indicators to monitor the implementation of the Digital Strategy.

People and culture

The expertise, skills and commitment of individual employees within an organisation are central to digital maturity on any level, whether the goal is transformation or introducing a new tool for using data. Core competencies often revolve around digital and data literacy, sometimes extending to advanced programming skills. Alongside these technical proficiencies, it is critical that employees understand the policies, processes and behaviours that promote the ethical use of data (OECD, 2020[14]). Furthermore, sector-specific knowledge and specialised expertise are also critical competencies, such as having sector-specific knowledge to understand the data landscape for risk analyses. Legal expertise is also valuable for navigating the legalities of data access, privacy, storage, and security. A digital-ready culture is not only about having the right set of skills and experiences available, but it demonstrates tangible ways that leadership and employees rally around digital goals. This can manifest in different ways, such as having policies that allow for the experimentation of new technologies or providing training for employees to improve their digital skills (OECD, 2022[4]).

Key practices in this dimension include:

- Ensure that leadership visibly endorses and partakes in digital initiatives, embodying a top-down commitment to the organisation's digital aspirations.
- Develop and implement a change management and continuous learning plan that focuses on enhancing digital and data literacy, as well as sector-specific knowledge.
- Introduce and encourage training programmes targeting technical proficiencies like advanced programming and data ethics.
- Institute clear policies that favour experimentation with new digital tools and technologies to foster innovation and a “trial-and-error” mentality.
- Establish guidelines on the ethical use of data, ensuring that staff understands and adheres to them.
- Prioritise and establish mechanisms for internal knowledge sharing, facilitating the dissemination of sector-specific, technical and legal expertise.
- Promote a culture of collaboration and digital empowerment, where employees at all levels feel engaged and invested in digital transformation objectives.
- Collaborate with legal experts to navigate the intricacies of data laws, ensuring the organisation remains compliant while maximising its digital potential.
- Implement feedback loops to understand employee challenges and needs in the digital landscape, adjusting strategies based on this feedback.
- Regularly evaluate the digital skills gap within the organisation and adjust training programmes accordingly.

Technology and processes

While technology, including IT systems, tools, and software, and the processes encompassing them are vital components of digital maturity, they should not be perceived as the primary objectives. The broader vision for digital transformation or the intent of any given project should inform technological advancements rather than being led by them. This underscores the importance of tailoring technology to specific needs. The term "state-of-the-art" is contextual, acknowledging that digital services, IT mechanisms, and tools differ in their complexity, resource needs, functionality, and alignment with various organisational goals. Given the rapid evolution of technology, expending resources without a lucid objective can lead to a waste of resources. Thus, a pragmatic approach involving cost-benefit analysis can guide judicious decision making about technological investments. This analysis can consider collaborative investments or

leveraging open-source technologies that may offer additional public benefits, such as the promotion of systemic transparency and collective technological development. Moreover, it is critical that considerations about adopting new technologies include assessments of their impacts on society, human rights and privacy, among other issues, to avoid exacerbating risks of discrimination and digital exclusion.

Key practices in this dimension include:

- Ensure any technology adoption aligns with the strategic objectives or specific goals of the organisation.
- Understand current capabilities, identify gaps, and ensure alignment with the organisation's digital maturity and objectives.
- Before investing in any new technology, gauge its potential return on investment and long-term sustainability.
- Start with a minimum viable product or proof-of-concept to test and validate new technologies or digital tools.
- Given the fast-paced nature of technological evolution, adapt and update tools and systems based on changing needs and feedback.
- Regularly research and explore new technological advancements that could replace legacy systems and enhance organisational effectiveness and efficiency.
- When selecting technologies, consider how easily they can be scaled or adapted to changing organisational needs or goals.
- Ensure that technologies are user-friendly, meet the needs of the organisation, and are accessible to all relevant stakeholders.
- Ensure that any new technology or process integrates robust cybersecurity protocols to safeguard organisational data.
- Define roles, responsibilities, and decision-making processes related to technology adoption and usage.
- Facilitate channels for sharing best practices, lessons learned, and feedback regarding technology tools and processes.
- Establish mechanisms to assess the social, human and ethical impact of adopting new technologies and mitigate associated risks, including those concerning data privacy, discrimination and digital exclusion.

Environment and partnerships

National frameworks, ranging from laws to directives, can considerably shape the trajectory of institutional digital transformation. For instance, a country's legal and policy framework can lay the foundation for managing digital governance, data stewardship and the sharing of data. These external parameters can influence the effectiveness of a digital transformation project, either limiting or propelling the adoption of digital technologies. Within an organisation, while robust data governance streamlines the sharing and accessibility of data, the true essence of data sharing transcends just infrastructure or processes. Cultivating collaborative relationships between entities, including industry, academia and civil society organisations, is an indispensable cornerstone for advancing digital maturity and ensuring that necessary safeguards are in place.

Key practices in this dimension include:

- Stay abreast of updates to knowledge of laws, policies, and guidance related to digital governance, data management, and sharing.
- Engage with policymakers to advocate for supportive laws, regulations and policies that bolster the goals of digital initiatives.
- Establish data-sharing protocols that align with both internal goals and external legal requirements, allowing for efficient and timely exchange of information.
- Define institutional roles, responsibilities and expectations for all digital initiatives that involve collaboration with external stakeholders.
- Ensure that partnerships are mutually beneficial, fostering a sense of shared ownership and collective achievement.
- Establish channels to gather feedback from partners, ensuring continuous improvement in collaborative endeavours.
- Encourage an organisational mindset that values partnerships as a key enabler of digital growth.
- Establish relationships with other organisations, both within and outside the government (e.g. industry, academia, civil society), to promote collaborative digital initiatives, share best practices, and ensure ethical use.

2 Generative AI: Challenges, risks and other considerations for integrity actors in government

Many of the opportunities and benefits that generative AI and LLMs offer, as discussed in Section 1, come with a unique set of challenges, risks and technical considerations. The 59 integrity actors that responded to the OECD's questionnaire highlighted issues that are relevant for their own context. However, many of the challenges and concerns they raised regarding the development, deployment and scaling of LLMs are relevant across different types of organisations and regions. This section explores these aspects of generative AI, particularly LLMs. It provides insights that can help integrity actors to understand and anticipate the range of challenges that this new area presents and be better positioned to overcome them if and when they arise.

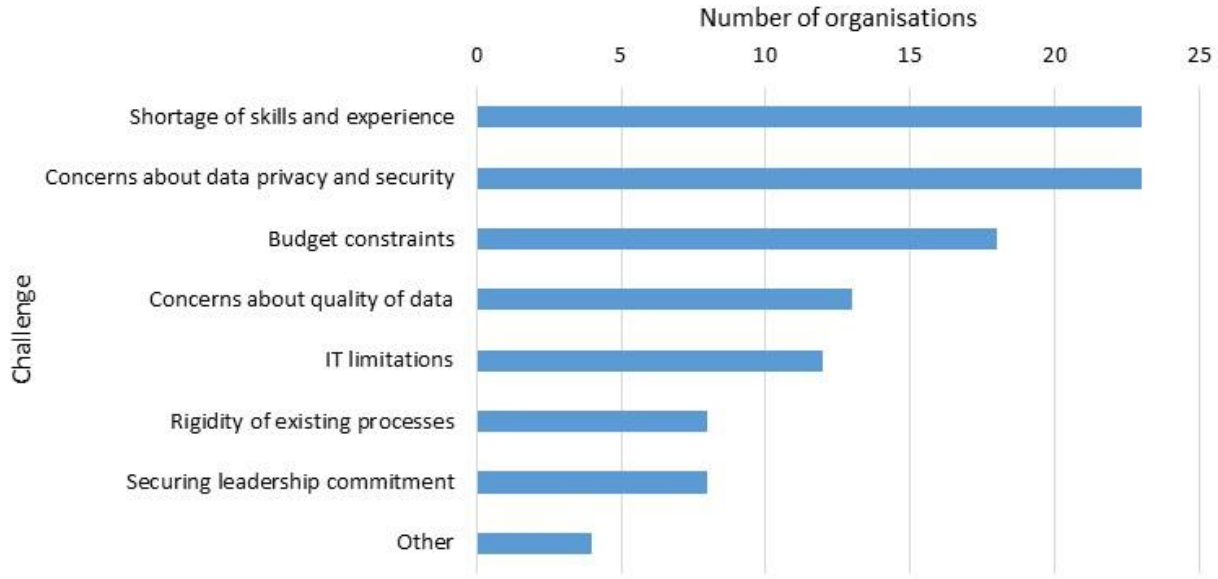
2.1. Overview of main challenges for integrity actors to adopt generative AI and LLMs

2.1.1. Integrity actors cited a shortage of skills and IT limitations as the biggest challenges they face to implement generative AI and LLMs

The OECD asked questions to understand the nature of challenges that integrity actors face to adopt generative AI and LLMs. Shortage of skills and experience ranked at the top of organisations' concerns, and this was the main challenge identified by anti-corruption agencies (see Figure 2.1). Organisations identified challenges related to preserving data privacy and security just as frequently; this issue was of particular concern to SAI respondents. Budget constraints, quality of data and IT limitations were also flagged as either the greatest or second greatest challenge by at least 10 of the organisations. Relatively fewer respondents highlighted concerns about the rigidity of existing processes or securing leadership commitment. One SAI noted that creating a business case for using and integrating generative AI into its operations is a challenge.

Figure 2.1. Main challenges for deploying generative AI and LLMs

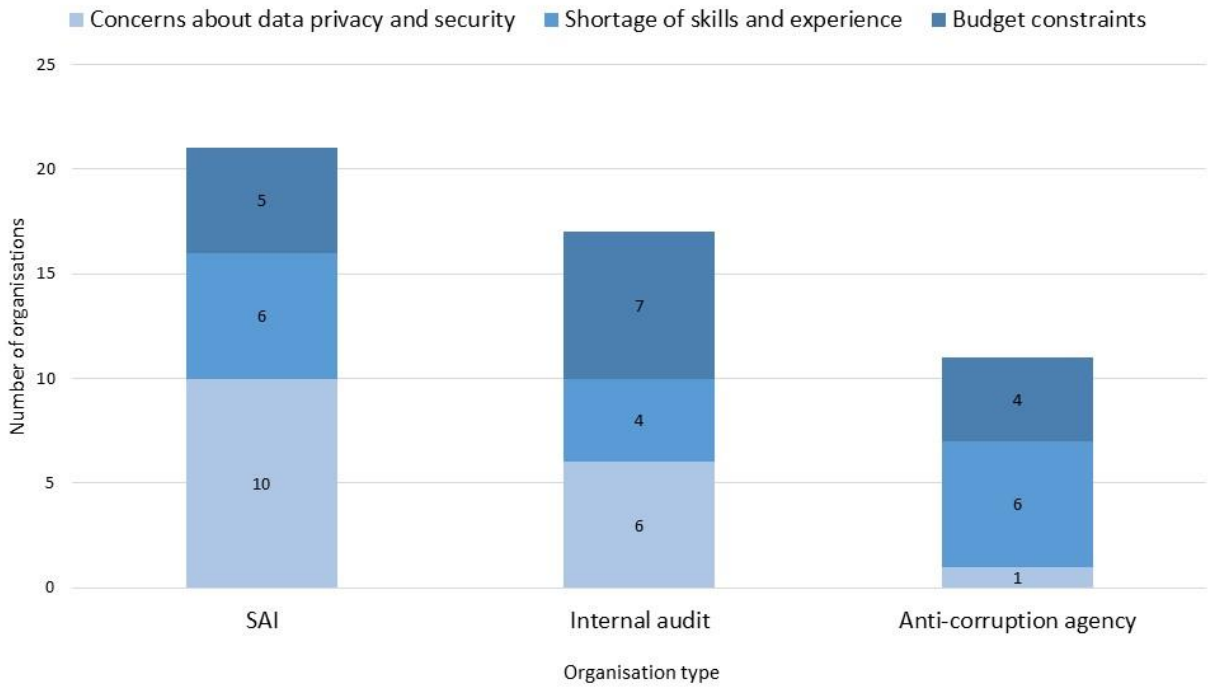
What are the biggest challenges your institution faces concerning the adoption of Gen AI and LLMs in general?



Note: "Number of organisations" refers to the number of organisations that selected each challenge as either their greatest or second greatest concern. Possible responses included the following: 1) Shortage of skills and expertise; 2) Concerns about data privacy and security; 3) Budget constraints; 4) Concerns about the quality of data inputs and outputs (e.g. biases and "hallucinations"); 5) IT limitations for developing and maintaining LLMs (e.g. IT systems and computing capacity); 6) Rigidity of existing structures or processes; 7) Securing leadership commitment and support; and 8) Other.
Source: OECD Questionnaire

Concerns about budget constraints were somewhat more pronounced among internal audit bodies relative to SAIs and ACAs, taking into account responses as a percentage of the total number of institutions by type (see Figure 2.2). Many institutions expressed that due to resource constraints, they either lack the sufficient financial, human, and technical resources needed to employ LLMs entirely or their staff does not have sufficient data literacy to make the use of such tools possible. One category of challenges the questionnaire did not clearly capture was that of methodological limitations. For instance, one ACA noted that its biggest challenge was having sufficient data to be able to develop an LLM. This points to a hierarchy of needs when it comes to developing LLMs. Given many institutions' early stage of development, in practice, some institutions appear to be focused on more technical challenges of developing viable proof-of-concepts, while recognising other challenges lie ahead (e.g. ensuring data privacy and security), particularly as they scale and roll-out LLMs.

Figure 2.2. Main challenges for deploying generative AI and LLMs by type of organisation



Source: OECD questionnaire

Tailored education is pivotal for overcoming challenges associated with skill and expertise deficits for using generative AI and mitigating associated risks. As illustrated by the experience of the European Court of Auditors (ECA), tailoring trainings involves adapting the curriculum to make courses available that are dedicated to generative AI and LLMs. Additionally, it means ensuring training content illustrates concrete uses cases and links generative AI tools to processes that are familiar to the trainees, which in the ECA's case would be auditors. Box 2.1 further describes the ECA's initiative to develop its trainings for generative AI.

Box 2.1. The generative AI training programmes of the European Court of Auditors (ECA)

The ECA is the supreme audit institution (SAI) of the European Union (EU) and is responsible for auditing the EU's finances as well as co-ordinating good practices across the SAIs of the 27 EU member states. The ECA is exploring how generative AI can be employed to make its audits more efficient and effective. As of February 2024, the ECA has developed two trainings on generative AI and is preparing several more. The trainings were developed in response to increasing demand among staff for guidance on how to employ generative AI tools in light of ChatGPT's growing popularity.

The ECA therefore developed an introductory training on generative AI that covers both how it works and ways in which existing generative AI tools can be used in auditing. It also offers advanced training where staff can develop their own machine-learning tools. The ECA has repeated the introductory training in response to high demand from staff, demonstrating that staff are eager to employ these tools once they have the proper knowledge.

A key distinguishing feature of the trainings offered by the ECA is their focus on integrating examples from existing audit work. The trainings explain how generative AI could have been used at different stages of past audits that staff are already familiar with, which promotes an understanding of the benefits and risks of generative AI on a practical basis, rather than a theoretical one. The training also teaches staff how to critically evaluate the outputs of generative AI.

The ECA is planning to develop more trainings based on areas of high demand and/or high risk. These include legal and copyright risks related to generative AI, conducting cybersecurity audits using AI, and including AI-based risks in IT audit methodologies. A training on prompt engineering in the context of generative AI is also under consideration.

Source: OECD interview with the European Court of Auditors

Concerning regional challenges highlighted in the questionnaire responses, several institutions in EU countries highlighted the need to ensure compliance with the General Data Protection Regulations (GDPR) and the EU's AI Act. The GDPR restricts the terms under which organisations in EU countries can reuse personal data, namely by requiring user consent. Understanding the full impact of the GDPR on AI or the integrity actors' use of it is beyond the scope of this paper. Nonetheless, respondents to the OECD's questionnaire highlighted this issue as a key consideration in their implementation of LLMs, which has resulted in them taking a more cautious and deliberate approach. As discussed later, in the context of AI, integrity actors may also face tensions between the need for algorithmic transparency and protecting data privacy.

The European Parliament approved the AI Act at the time of writing this paper in March 2024, with a formal endorsement by the Council of the EU needed before it enters into force. The Act establishes obligations for AI developers and users based on potential risks and the level of impact of the AI system, with the aim of protecting fundamental rights, democracy, the rule of law and the environment from high-risk forms of AI (European Parliament, 2024[15]). The effect of the AI Act on government entities as users of AI, including integrity actors relying on foundation LLMs of private companies, remains to be seen as EU countries turn towards implementation and enforcement of the Act. Section 2.3 below explores themes related to the purpose of the EU's AI Act, including challenges and considerations for integrity actors concerning the promotion of trustworthy AI and responsible use of LLMs.

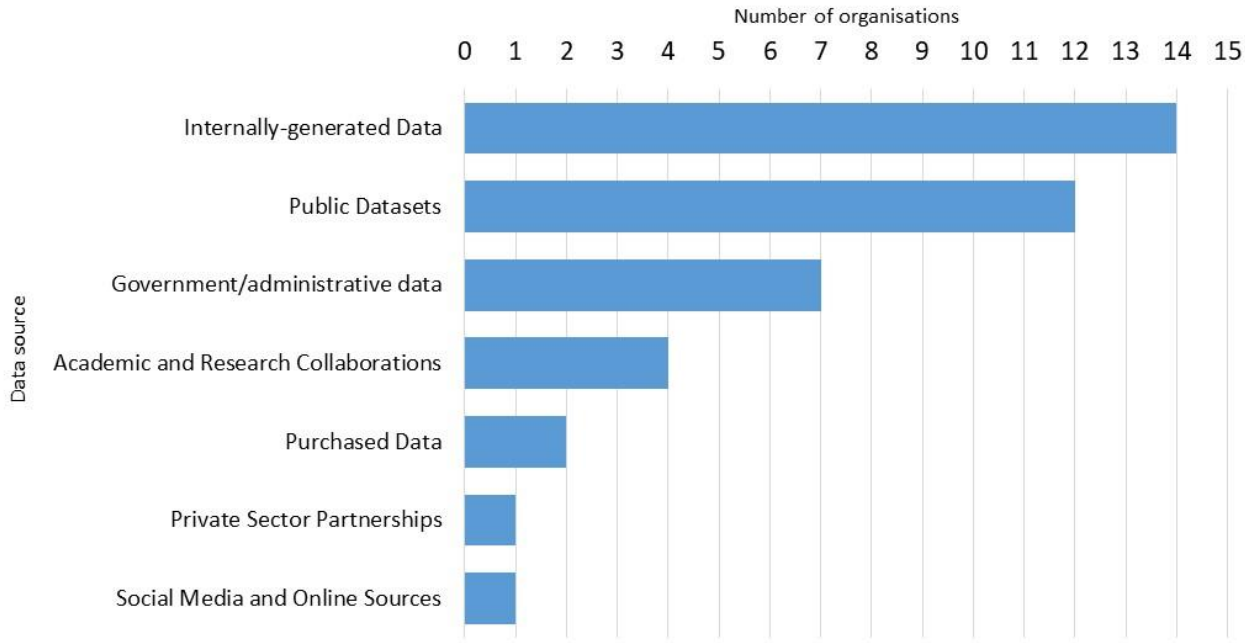
2.1.2. Advice for piloting LLMs includes first incorporating generative AI into low-risk processes and considering the requirements for scaling early on

The questionnaire asked specifically about challenges related to piloting and scaling generative AI initiatives, focusing on the 17 respondents that are either experimenting with or integrating generative AI tools. Many of the main challenges echoed those highlighted in the figure above, including shortage of skills, IT limitations and concern about the quality of data inputs and outputs (e.g. biases and hallucinations). Only one institution expressed concerns about securing leadership commitment, although respondents may have been less likely to select this option if superiors monitored their responses. However, since few institutions have reached this level of maturity and there were therefore fewer responses to this question, the spread in responses was fairly even. Several themes came to the forefront in the responses to questions about piloting and scaling generative AI initiatives:

- Start by incorporating generative AI into low-risk areas and processes. Many of the institutions that have reached the piloting stage seem to be focusing on incorporating generative AI, with a focus on LLMs, into relatively low-risk processes, such as document querying, writing document summaries and press releases, and answering user questions. Such an approach can help build capacity in areas where mistakes are less costly—either financially or from a compliance perspective—before they scale LLMs to riskier and more analytical tasks, including those that require more financial resources.
- Consider the IT requirements not only for piloting, but for scaling as well. When piloting LLMs, it is first necessary to establish certain prerequisites for IT infrastructure. This includes computational and storage resources, including the availability of high-performance computing power, data storage, and data management capabilities. Over half of the 17 respondents with LLM initiatives ranked this as the number one IT challenge, followed by challenges related to software tools as well as system scalability and integration.[5] One respondent noted that having the rights tools in place first is just as important as having the right algorithms.
- Consider internally-generated data to demonstrate value and establish quick wins. This data could be internally held and/or produced by the government body itself or come from another source. The integrity actors that responded to the OECD's questionnaire are primarily relying on internally-generated data or public open datasets, potentially because they view this approach as a lower risk than using other data sources (see the next section for a discussion on Retrieval-Augmented Generation). Comparatively fewer organisations are using other government data, while only a limited number of organisations are using data purchased from the private sector, obtained through a public-private or academic partnership, or obtained from social media or another online source (see Figure 2.3).

Figure 2.3. Primary data sources for building LLMs among questionnaire respondents

From which sources does your institution primarily acquire the data used for building and training your LLM(s)?



Note: Possible responses included the following: 1) Internally Generated Data: Data generated from within our own institution (e.g. reports, administrative records, etc.); 2) Public Datasets: Data sourced from publicly available datasets (e.g. government open data portals, public research datasets). 3) Government/administrative data: Data sources produced or owned by government entities, but are not public or open. 4) Private Sector Partnerships: Data obtained through partnerships or agreements with private sector entities. 5) Purchased Data: Data procured from commercial data providers or brokers. 5) Academic and Research Collaborations: Data obtained through collaborations with academic or research institutions. 7) Social Media and Online Sources: Data extracted from social media platforms, websites, and other online sources; 8) Other.
Source: OECD questionnaire

One notable difference between the challenges identified for piloting LLMs versus scaling them is the emphasis on both data privacy and security as well as budget constraints. In the initial phases of testing LLMs, the primary challenges highlighted involve data privacy and security, alongside concerns about data quality. Out of 17 organisations, only two mentioned budget constraints as an issue during this pilot phase. Conversely, when it comes to expanding the use of generative AI and LLMs, budget constraints emerge as a more significant challenge, with fewer organisations expressing concerns about data privacy and security at this stage. This may reflect an evolution in maturity in terms of managing data privacy and security issues, as well as the increased resource needs when scaling LLMs that are not present early on.

2.2. Building a generative AI and LLM capacity within institutions responsible for integrity and anti-corruption

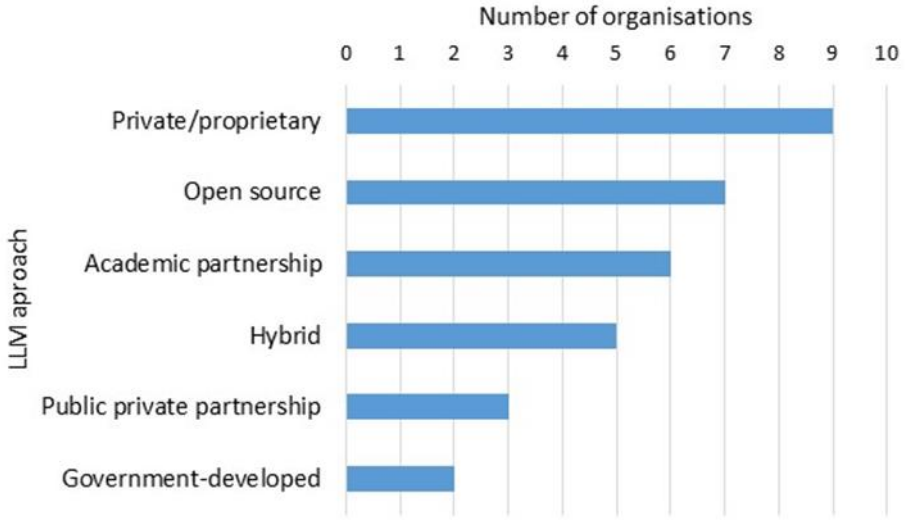
2.2.1. Integrity actors mostly rely on turnkey foundation LLMs developed by technology companies

Integrity actors have multiple pathways for piloting and scaling LLMs. These include leveraging open-source LLMs; utilising models developed by private companies for their advanced capabilities; or embarking on their own development projects. Collaborative efforts with the private sector or academia, as well as hybrid approaches that combine these elements, present viable alternatives. Of these options, open-source LLMs offer some algorithmic transparency. However, due to intellectual property restrictions, it can still be difficult for users to review their source code and training data, thus limiting the degree of transparency. Other mechanisms that promote interpretability, explainability, such as user-friendly explanations of decisions made, can help promote transparency in decision making internally and to the public at large (see Section 2.3).

The integrity actors that responded to the OECD's questionnaire predominantly use open-source and private sector models, which according to several respondents, helped to overcome constraints in financial and human resources (see Figure 2.4). Several of the integrity actors that reached the development stage of using LLMs (see Section 1) highlighted the use of multiple approaches to testing or using them.

Figure 2.4. Integrity actors' approach for using LLMs

What is the general approach your institution is taking to test and/or use LLM(s) for your operations?



Note: This question was only asked to questionnaire respondents who have reached the stage of developing generative AI models. Possible responses included the following: 1) Open Source Model: We use or develop LLMs based on open-source platforms or technologies. 2) Private/Proprietary Model: We use LLMs developed by private companies (e.g. ChatGPT by OpenAI). 3) Hybrid Model: We use a combination of open-source and private/proprietary LLMs. 4) Public-Private Partnership: Our LLMs are developed or used in collaboration with private entities under a public-private partnership model. 5) Government Developed and Maintained: Our LLMs are exclusively developed and maintained using government resources without private sector involvement. 6) Research and Academic Collaboration: We are engaged in collaborations with academic or research institutions for the development or use of LLMs.
Source: OECD questionnaire.

Integrity actors are leveraging a variety of LLMs to enhance their operations, most prominently models developed by companies like OpenAI, Google and Meta. Notable LLMs include OpenAI's Generative Pre-trained Transformer 4 (GPT-4), Google's Pathways Language Model (PaLM), and Meta's Open Pre-trained Transformer (OPT-175B), alongside other models like Google's BERT and Meta's LLM, Meta AI (LLaMA) (OECD, 2023[1]). These models offer foundational capabilities that can be specifically tailored to the unique requirements of integrity actors through techniques like Retrieval-Augmented Generation (RAG), which enriches LLMs with information from additional databases, including their own data sources. See Box 2.2 for further explanation of RAG.

Box 2.2. Retrieval-Augmented Generation for LLMs

Retrieval-Augmented Generation (RAG) is a technique developed to improve how large language models (LLMs), like the ones behind chatbots and virtual assistants, handle information. For different reasons, including reliance on old data, LLMs can provide incorrect answers and it can be difficult to understand how they derived at a particular response. RAG can help to address these challenges by allowing LLMs to access additional databases that can keep information current, which is particularly useful when applied to specialised domains or knowledge areas. For integrity actors, RAG can be an effective means for fencing-in their internal data sources, while improving the accuracy, relevancy and trustworthiness of a model's output.

RAG begins with identifying pertinent documentation and extracting vital text from it. Then, it breaks this text down into smaller parts and transforms these parts into a format (i.e. embeddings) that the model can understand and store efficiently. These pieces of information are kept in a special database (i.e. vector databases). When someone asks the model a question, it can look through this database to find up-to-date and accurate information to add to what it already knows before giving an answer.

For situations where it is critical for a model to provide facts that are current and accurate, such as when dealing with confidential information or needing to keep a clear record of data sources, the U.K.'s Generative AI Framework recommends using RAG. This approach can help to ensure that the model's answers are based on reliable data, making it particularly valuable for organisations focused on maintaining high levels of accuracy and accountability.

Source: (UK Government, 2024[16]; Gao et al., 2023[17])

On a technical level, integrity actors that responded to the questionnaire are either using an existing turnkey model—a model which is available in a ready to use form—without fine-tuning (7 out of 17 respondents), or they are fine-tuning a foundation model (7 out of 17 respondents). They primarily deploy GPT-4 and its predecessor, GPT-3.5, alongside BERT and LLaMA-2 for their advanced text processing needs. Several integrity actors employ platforms like ChatGPT in their generic form for broader tasks. However, the dependency on commercial LLMs poses challenges, particularly in data usage transparency and the risk of biases, as explored below (OECD, 2023[1]). Several integrity actors highlighted the use of RAG to fine-tune models. For instance, in the questionnaire responses, several SAIs highlighted the use of RAG for incorporating their own repositories of data and documents into the model, thereby further enhancing the customisation of the LLM.

A few respondents highlighted the use of LLMs tailored to the national context, such as in Norway and France, where integrity actors are making use of bespoke open-source tools. For instance, Box 2.3 provides an example of how an entity in the French government fine-tuned Llama to create a tool aimed at improving the efficiency and efficacy of parliamentary sessions by generating summaries of legislative proposals. This tool offers inspiration in a number of areas. Oversight bodies could benefit from an LLM that summarises complex legislative texts into concise versions to gain a quicker understanding of issues

that are relevant for audit engagements and decision making. In addition, ACAs could use a similar approach to create summaries to detect risks of undue influence in legislative proposals, such as clauses that might be overly beneficial to a specific group without sufficient justification.

Box 2.3. France's LLaMandement for summarising legislative text

Creating concise summaries is crucial in managing the legislative process, where tens of thousands of amendments, each spanning roughly two pages, are processed annually. These summaries are vital for a wide range of stakeholders—government officials, ministers, commission members, deputies, senators, administrative agents, journalists, and citizens—to quickly understand and discuss amendment contents without revisiting the full texts. AI-supported tools, especially those using LLMs, can play a significant role in this context. In contrast to other techniques, LLMs have the potential to efficiently distil vast amounts of complex legal texts into easily understandable information, enhancing efficient communication and informed decision making.

Recognising this opportunity, the Digital Transformation Delegation of the French Directorate General of Public Finances launched the "LLaMandement" project to automate the handling of legislative amendments. This project uses LLMs to assign amendments to the appropriate ministerial departments, search for past similar cases, and synthesise amendments into clear, ideally neutral summaries. The tool is designed to enhance the efficiency and accuracy of administrative work, supporting individuals to analyse bills and process amendments, especially during peak legislative periods.

LLaMandement draws on data from the Inter-ministerial Digital Management System for Legislative Amendments (SIGNALE), and it uses ministers' bench memoranda for training the model to ensure comprehensive understanding across different ministerial contexts. The developers of LLaMandement were sensitive to the possibility that the model would create biased results or promote misinformation. To address this concern, they used the Bias in Open-ended Language Generation Dataset (BOLD), a dataset used for evaluating biases in LLMs, particularly in open-ended text generation. Using BOLD, the developers assessed LLaMandement for biases related to gender, ethnicity and political ideology. They concluded the model reliably exhibited very few errors and the results were unbiased and neutral for different groups of people and beliefs.

Source: (Gesnouin et al., 2024[18])

2.2.2. Overcoming language barriers inherent in using or fine-tuning many off-the-shelf LLMs is a key challenge for integrity actors

One challenge that many organisations highlighted was the lack of existing LLMs trained in their native language. A 2023 study found that 38% of NLP models, which include LLMs, on the open-source platform Hugging Face are trained in English, followed by Spanish, German, and French (all at around 5%) (OECD, 2023[1]). Very few LLMs are trained in languages other than English. As illustrated in feedback from integrity actors who participated in OECD workshops and who responded to the questionnaire, integrity actors have had to invest extra time and energy into training their LLMs in their national language(s), usually by feeding the model regulations and reports written in native languages. This issue undermines the ability of many institutions to rely on existing LLMs with limited fine-tuning, as most LLMs are trained in English and a handful of other common languages (e.g. Spanish).

To enhance linguistic accuracy in local contexts, some countries are investing in the development of native-language LLMs that will be open source. Examples include the Netherlands' GPT-NL and Sweden's GPT-SW3, which are designed to excel in processing national languages by training on locally relevant texts.

These initiatives not only reduce dependency on technology companies but also offer improved performance in handling sensitive integrity-related data. Box 2.4 highlights the approach of the Office of the Comptroller General (*Controladoria Geral da União*, CGU) of Brazil to overcoming this and other challenges it faced while piloting its own LLMs.

Box 2.4. The Office of the Comptroller General (CGU) of Brazil's approach to piloting LLMs

CGU is an anti-corruption body within the public administration that is responsible both for financial management and transparency measures. It plans to use generative AI to support a variety of tasks, including inference of risks from internal audit reports, analysis of management response to internal audit recommendations, drafting of audit engagement findings, responding to support requests related to the asset and conflict-of-interest declaration system, and querying internal audit reports. The institution does not foresee generative AI replacing auditors but rather as a "co-pilot" that can help improve their efficiency. To this end, CGU's Data Intelligence Unit has invested in fine-tuning Llama-2 into their own LLM called Llama-2 GOV BR.

CGU encountered several challenges in its attempts to incorporate generative AI in its work. These included challenges related to inference time, scalability, costs, data sensitivity, and the content policy. CGU found that one way to overcome several of these challenges was by investing in a comparatively smaller LLM. Such models can achieve similar performance to larger models if trained well to do specific tasks with the added benefits that they can be served by local infrastructure, which reduces costs and inference time, improves scalability, keeps sensitive data on the organisation's premises, and allows for local management of the content policy.

When considering how best to deploy generative AI in their institution CGU also encountered the problem that existing LLMs did not perform well in Portuguese. However, since developing a new LLM from scratch is extremely expensive, it opted to fine-tune the Llama-2 model for its purposes. By pre-training the model with 10 million lines of high-quality Portuguese text from sources including audit reports, federal legislation, and PhD theses, CGU was able to reach a point where its LLM performed well enough to be used in its work. The CGU has plans to develop further monitoring and evaluation activities to ensure the LLM's reliability before rolling it out for day-to-day use.

Source: Meeting of OECD's Community of Practice on Technology and Analytics for Public Integrity: "Generative AI for promoting integrity and accountability in the public sector" (8 November 2023)

2.3. Ensuring the responsible development and use of generative AI and LLMs by integrity actors

2.3.1. Integrity actors recognised the need for safeguards, but more can be done to ensure the responsible and ethical use of AI as initiatives mature

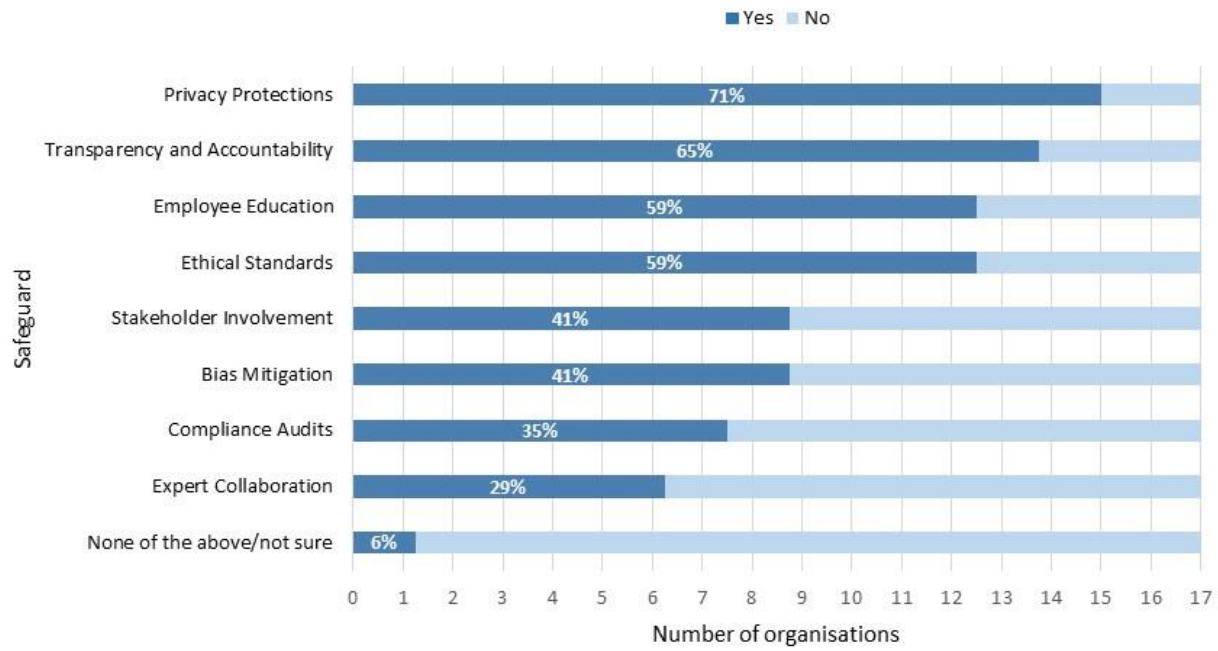
When asked to select from a range of challenges concerning the development and implementation of generative AI, including LLMs, and almost all integrity actors with relevant initiatives ranked issues surrounding compliance and ethics at the bottom of the list. This challenge involves navigating ethical, legal and privacy concerns, as well as regulatory compliance. Other challenges, such as technical development, resource management, and data management, ranked the highest (in that order) in terms of integrity actors' priorities for developing and implementing LLMs. This may reflect the current maturity of these integrity actors, most of which remain in the early stages of incubating ideas or ad hoc experimentation. Nonetheless, these challenges are important to consider during the design phase for the

reasons discussed below. Challenges will likely become more acute as public institutions solve more technical issues and are using LLMs more frequently and for more advanced tasks. Moreover, as regulations develop in this area it will also put pressure on these institutions to put measures in place to ensure compliance.

Integrity actors also offered insights about the measures their organisations employ to ensure the responsible use of AI as well as LLMs. Of the 17 organisations with LLM initiatives, the majority are employing the following measures: privacy protections, transparency and accountability measures, employee education and ethical standards. For instance, privacy protections include safeguards to adhere to data protection standards, and transparency and accountability broadly refers to measures to ensure open decision making about the use of AI, including redress mechanisms for citizens (OECD, 2022[19]) (see the note in Figure 2.5 for further explanation about safeguards).

Figure 2.5. Safeguards to ensure responsible use of AI and LLMs

What measures does your institution employ to ensure responsible AI and LLM usage?



Note: Possible responses included the following: 1) Ethical Standards: Implementation of ethical guidelines and policies; 2) Transparency and Accountability: Ensuring open AI decision making and maintaining accountability; 3) Bias Mitigation: Actively addressing biases to promote fairness; 4) Privacy Protections: Adhering to privacy and data protection standards; 5) Compliance Audits: Conducting regular ethical and legal compliance assessments; 6) Stakeholder Involvement: Engaging with stakeholders for input and addressing concerns; 7) Employee Education: Offering training and awareness programmes on responsible AI; 8) Expert Collaboration: Working with external experts for ethical and legal guidance; 9) None of the above/not sure; 10) Other. None of the respondents selected other.
Source: OECD questionnaire

Fewer institutions identified measures in place for bias mitigation, stakeholder involvement, compliance audits or expert collaboration as mechanisms to ensure responsible use of LLMs. The ranking of bias mitigation is notable. While it is unclear whether surveyed institutions do not see this as an issue or do not know what measures they should put in place, the issue of bias and hallucinations is a critical area for concern as LLMs become increasingly mainstreamed. Not only does this issue present policy, regulatory

and technical challenges, but it also poses political and reputational risks for those organisations that are experimenting with generative AI.

As integrity actors develop generative AI tools, including LLMs, they will need to contend with the issue of bias. Sampling bias is one form of bias that can be difficult to detect. This type of bias occurs when the data that underly a model are not actually representative of the population that they are meant to represent (Berryhill et al., 2019[20]). Sampling bias can be further broken down into historical bias related to pre-existing patterns in training data, representation bias arising from missing variables or an inadequate sample size, and measurement bias related the erroneous omission or inclusion of certain variables. For example, AI models designed to assign a corruption score to specific individuals based on previous conviction data could reflect biases related to higher wrongful conviction rates for racial minorities (Köbis, Starke and Rahwan, 2021[21]), thereby perpetuating the discrimination, marginalisation or exclusion of large segments of the population.

Similarly, when training algorithms, there is also the possibility of statistical bias. Statistical bias occurs when a model consistently makes the same error in prediction based on the expected outcome (Berryhill et al., 2019[20]). It is comparatively easy to detect. If a model consistently overestimates a value by the same amount, for example, the model simply requires more fine-tuning. This is fundamentally a problem with the model itself that those training it will need to resolve, and therefore may be less relevant for integrity actors that rely on LLMs of private companies and have less control over the design of the model.

Furthermore, if an LLM is trained disproportionately on texts produced by—or reflecting the experience of—certain categories of individuals, the LLM may eventually display more favourable views towards these categories of individuals or more unfavourable views towards other categories of individuals. Measures to mitigate this type of bias can include taking stock of training data for underrepresented groups, curation or semi-automatic curation of datasets to reach fairer results, as well as explainability and interpretability research and applying auditing processes (Lorenz, Perset and Berryhill, 2023[3]). In general, including more parameters when training a model reduces bias, but this can have other negative effects, such as increasing energy requirements or infringing more on personal privacy (OECD, 2023[1]), so integrity actors should carefully weigh these concerns when training LLMs. Another innovative approach is “red teaming” whereby researchers use one LLM to identify biases in another (Lorenz, Perset and Berryhill, 2023[3]).

For bias mitigation measures to be successful, there must first be a recognition of both the threat and consequences of biases. However, existing research on initiatives for AI as an anti-corruption tool uncovered a general lack of concern about bias mitigation, as well as a lack of accountability and transparency mechanisms to ensure the necessary bias mitigation was taking place (Odilla, 2023[22]). Examples that illustrate potential consequences of ignoring such issues can be found in different countries and sectors. For instance, the “Toeslagenaffaire” was a child benefits scandal in the Netherlands where the use of an algorithm resulted in tens of thousands of often-vulnerable families being wrongfully accused of fraud, as well as hundreds of children being separated from their families. This extreme case led to the collapse of the government. In Australia, in what became known as the “Robodebt scheme,” a data-matching algorithm calculated overpayments to welfare recipients that resulted in 470 000 incorrect debt notices and the sending of EUR 775 million in undue debt payments by welfare recipients, leading to a national scandal and a Royal Commission (OECD, 2023[23]).

While these examples are about AI in general, going beyond generative AI or LLMs, they illustrate the potential for severe political and social consequences that are relevant for integrity actors to consider as they embark on the use of generative AI and LLMs. It is critical for integrity actors to ensure they are taking the necessary steps to mitigate bias—including by ensuring compliance with national non-discrimination legislation—and sufficiently documenting how they have done so to establish trust in the tools that they have developed. For more advanced use cases, when appropriate, integrity actors can also promote redress mechanisms for citizens affected by algorithm-driven decisions (OECD, 2022[19]). Finally, maintaining a focus on a human-in-the-loop system, whereby trained humans play a central role in the

development of models and creating a continuous feedback loop, can help to mitigate the risk of machine biases that manifest into harmful decisions and actions.

The use of AI in general, whether generative AI or other forms of AI, to promote integrity and combat corruption poses a unique set of ethical concerns. Integrity actors already process large amounts of personal data, such as in mandatory interest or asset declarations. This means they must be careful that any use of generative AI to process this data protects individuals' privacy and that the entity does not disseminate data that would not otherwise be publicly available. For more detail on how the Corruption Prevention Commission (*ՀՀ Կոռուպցիայի կանխարգելման հանձնաժողով*, CPC) of Armenia has worked to address ethical concerns in this area see Box 2.5. Ethical issues can also arise when working with crowdsourced data, such as whistleblower complaints. AI models may have difficulty distinguishing founded complaints from unfounded ones, which could lead to wrongful denunciation of public officials or wrongful decisions on cases concerning citizens. Research has shown that individuals are generally against algorithms making ethical decisions (Köbis, Starke and Rahwan, 2021[21]), which therefore requires that AI and generative AI models utilised in integrity bodies still have some level of human oversight.

Box 2.5. The Corruption Prevention Commission (CPC) of Armenia's use of AI to verify asset declarations

Armenia established the CPC in 2019 as part of a wider package of anti-corruption reforms, and upon its creation it assumed the responsibility for overseeing the electronic register of asset declarations. However, given the large number of officials required to submit these declarations and the CPC's limited resources, it was difficult to perform any meaningful checks of the content of these declarations. Initially, the electronic submissions were not even machine readable. The CPC therefore decided to build a data platform that would provide the necessary structure for data analysis and link to the databases of other state bodies. The CPC recently introduced an automated verification system that conducts an initial screening of asset declarations and identifies red flags. It is currently piloting the introduction of an AI component that would enable this system to learn from this process and identify new patterns in corrupt behaviour.

CPC noted that stakeholder engagement played a key role in developing a tool that would handle this sensitive data properly. Consultations with private sector actors both domestically and internationally helped the CPC gain a better understanding of the technical infrastructure that would be necessary for this tool to function and effectively and responsibly. Ultimately, these consultations led the CPC to invest in improving the quality of the underlying data first before experimenting with machine learning algorithms.

In other areas, different ethical considerations conflicted with each other, making finding a solution more difficult. For example, the desire to promote transparency by publishing the algorithm used to verify the asset declarations conflicted with the need to respect the privacy of those declaring, particularly given that Armenia has aligned its legal framework with the GDPR. In the end, the CPC determined that while publishing the asset declarations themselves was in the public interest and therefore permissible under the GDPR, publishing the algorithm was not.

It nonetheless remains important to be transparent about how the system is flagging declarations to maintain public trust, and the need to balance transparency and privacy will persist as development of this tool continues. In neighbouring Georgia, a lack of transparency about how the government is using AI has been undermining trust in the public institutions using these tools.

Source: (Izdebski, Turashvili and Harutyunyan, 2023[24])

2.3.2. Integrity actors can put a greater emphasis on monitoring and evaluating their AI activities, including consideration of model interpretability

High-level principles, standards and national normative frameworks offer a starting point for integrity actors to ensure they a prioritising the responsible use of AI as their initiatives mature. The OECD Recommendation on Artificial Intelligence (OECD, 2023[25]) identifies five value-based principles for the responsible use of AI (see Box 2.6). Many countries have adopted similar principles within their national frameworks for regulating AI. For example, Switzerland's Guidelines on Artificial Intelligence for the Confederation mirror the OECD Principles and add principles on regulatory compliance, stakeholder engagement, and actively shaping global AI governance (Federal Council of Switzerland, 2020[26]). They also contain more detail on complying with specific legal principles. The same is true of the Government-Wide Vision on Generative AI of the Netherlands, which outlines six specific areas of action to support the principles (Government of the Netherlands, 2024[6]). The Netherlands has also taken the approach of establishing a Government AI Validation Team to review pilot projects and ensure compliance with the principles, which can help mitigate risks related to irresponsible use. In Denmark, the Agency for Digitalisation (*Digitaliseringsstyrelsen*) has issued targeted guidelines for managers in public authorities to ensure responsible use of generative AI in their institutions (Danish Agency for Digitalisation, 2024[27]). Such ethical frameworks for AI generally also play an important role in mitigating risks related to generative AI specifically (Lorenz, Perset and Berryhill, 2023[3]). Integrity actors can consider these broad responsible use issues when developing generative AI tools and put the necessary safeguards in place to ensure responsible use.

Box 2.6. The OECD Principles on Artificial Intelligence

The OECD Principles on Artificial Intelligence, which are laid out in the OECD Council Recommendation on Artificial Intelligence, are divided into values-based principles and recommendations for policymakers. The five value-based principles that aim to encourage responsible use of AI in line with key values of OECD member states are as follows:

- Inclusive growth, sustainable development and well-being. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
- Human-centred values and fairness. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
- Transparency and explainability. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
  - to foster a general understanding of AI systems,
  - to make stakeholders aware of their interactions with AI systems, including in the workplace,
  - to enable those affected by an AI system to understand the outcome, and,

    - to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
- Robustness, security and safety. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
- Accountability. AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

When exploring ways of incorporating generative AI into their work, integrity actors should therefore make sure they are adhering to these principles. Given the sensitive nature of the data that these organisations hold and process, respecting human rights, transparency, and security are particularly important. The OECD Council Recommendation on Artificial Intelligence is under revision and is expected for adoption at the Ministerial Council Meeting in May 2024.

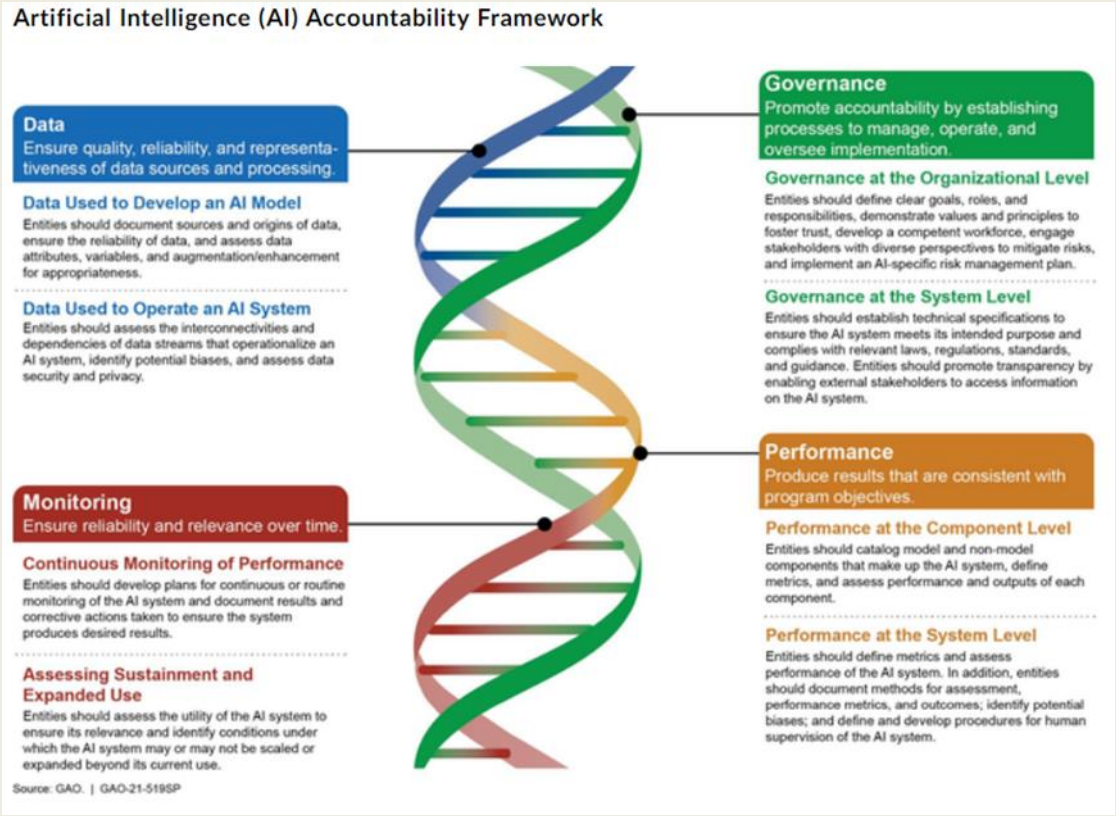
Source: (OECD, 2023[25]) (OECD, 2019[28])

Monitoring and evaluation play an important role in ensuring that AI use is indeed responsible. Monitoring during the implementation stage is necessary to ensure that risks are being mitigated and unintended consequences are identified. Public institutions can also take a risk-based approach to monitoring that involves higher scrutiny for processes with the potential for more severe negative consequences (Berryhill et al., 2019[20]), such as those related to anti-corruption. Box 2.7 provides more details on good practices for monitoring and evaluation within the US Government Accountability Office (GAO)'s AI Accountability Framework. Additional examples of how integrity actors, including SAIs and other oversight bodies, are promoting algorithmic can be seen around the world. This includes the development of a *General Instruction on Algorithmic Transparency* for public entities developed by the Chilean Transparency Council as well as a cross-border collaboration between the SAIs of Finland, Germany, the Netherlands Norway and the UK to develop a white paper on auditing machine learning algorithms (OECD, 2023[23]).

Box 2.7. The AI Accountability Framework of the US Government Accountability Office (GAO)

In order to ensure effective use of AI, GAO has developed a framework to evaluate the performance of AI systems to make sure they deliver value and remain fit for purpose over time. Pillars 3 and 4 of the framework on performance and monitoring, respectively, provide examples of best practices for monitoring and evaluation of data-driven tools (see Figure 2.6 below). These procedures for monitoring and evaluation are not only useful in guiding entities in their use of AI, but they could also be applied to other data-driven tools and systems, including those that do not employ AI.

Figure 2.6. GAO's Artificial Intelligence Accountability Framework



Regarding performance, at both the component level and the system level, AI models should be documented and assessed against predetermined performance metrics that are precise, consistent, and reproducible. Documentation should aim to address the following groups of questions:

- How are components and models solving defined problems? What is their intended use?
- How are the specifications and parameters are selected, evaluated, and optimised?
- How suitable are components and models to available data and operating conditions?
- How are components and models tested and what are the results?
- What ethical considerations exist? What biases and unintended consequences have been identified?
- What degree of human supervision is required and how was this determined?

When it comes to model evaluation, the selected performance metrics should be accurate and useful, and the justification for their selection and the person(s) responsible for their development should also be documented.

Once AI systems are put in place, a plan for continuous or routine monitoring should be developed. This helps ensure that AI models remain reliable and relevant. The plan should define acceptable levels of data and model drift that are based on a risk assessment and require documentation of monitoring activities and any corrective actions taken. As part of the monitoring of AI systems, their continued utility and any potential opportunities for scaling should also be assessed. Any decisions to retire or scale models or systems should be based on predefined performance metrics, and any updates that take place and their impact should be documented. Finally, throughout the process of monitoring and evaluation, it is important that entities keep in mind AI systems' consistency with their objectives and values in order to foster and maintain public trust.

Source: (US Government Accountability Office, 2021[29])

From a methodological perspective, one of the main challenges integrity actors face, like other organisations, relates to the limitations of LLMs in terms of interpretability, explainability and transparency. The breadth and variety of data that feed into LLMs, which are fundamental to their usefulness, present major challenges in tracing the connection between outputs and inputs. The complexity of the underlying architecture and decision-making mechanisms exacerbate this challenge (Shabsigh and Boukherouaa, 2023[2]) and can make it more difficult for citizens to understand how their government is making decisions or make appeals to protect their own rights and interests. In the integrity context, overcoming this challenge can be the difference between limited use (e.g. LLMs for summarising text) and more extensive integration of LLMs across core audit or investigative processes. For instance, preserving an audit or investigative trail, or providing justification for prioritising risks, are core tenets of the work of anti-corruption and oversight bodies alike. Their legal obligations and reputations rely on understanding the provenance of data that informs key decisions.

Challenges concerning interpretability, explainability and transparency of LLMs further highlight the importance of auditors, investigators and analysts maintaining professional scepticism and ensuring human-centred checks remain throughout the training and deployment of LLMs. There are no easy solutions to address this challenge. Government agencies have explored the use of decision trees to help illustrate the link between the results from AI systems and an explanation of how they came about (Berryhill et al., 2019[20]), and they have issued explainable AI toolkits to help assist in this area.[6] Academia offers additional ideas and insights. For instance, a group of researchers introduced a taxonomy of explainability techniques for LLMs, as well as metrics for evaluating generated explanations to improve model performance (Zhao, 2023[30]). Other research explores transparency within the unique context of LLMs and poses priorities and questions that can be helpful for integrity actors as they assess their own LLMs (see Box 2.8).

Box 2.8. Human-centred considerations for promoting transparency when evaluating LLMs

Integrity actors in government have a wide range of internal and external stakeholders to account for when considering the interpretability, explainability and transparency of the LLMs they develop. These stakeholders have different needs in terms of the what, when, and how of an AI initiative, given their different roles, responsibilities and levels of technical expertise in what researchers call the LLM ecosystem. Transparency in this context implies that relevant stakeholders can "form an appropriate understanding of a model or system's capabilities, limitations, how it works, and how to use or control its outputs."

There are several key areas and questions for consideration that are broadly applicable to any organisation that is developing LLMs, but they are especially useful for integrity actors in government given their responsibilities to the general public and other stakeholders. These areas and questions draw inspiration from the machine learning and human-computer interaction literature: and can provide integrity actors with a starting point for thinking about their own measures to promote transparency in the context of LLMs:

1. **Model reporting**
   - What information is needed to characterise the functional behaviour of an LLM?
   - What do different (and new) types of stakeholders need from model reporting frameworks?
   - What is needed beyond static documentation?
2. **Publishing evaluation results**
   - Who is the evaluation targeted at and for what purpose?
   - At what level should the evaluation take place?
   - How should LLM limitations and risks be evaluated?
3. **Providing explanations**
   - How can the organisation provide faithful explanations for the LLM, knowing that it is the ultimate black box?
   - What explanations are appropriate for LLM-infused applications?
4. **Communicating uncertainty**
   - What is a useful notion of uncertainty for LLMs?
   - What are the most effective ways to communicate uncertainty?

These questions are meant to guide future research, but they can also provide integrity actors with a starting point for thinking about their own measures to promote transparency in the context of LLMs, as well as ways to strengthen evaluation mechanisms with an approach that is tailored to the unique LLM context.

Source: (Liao and Vaughan, 2023[31])

2.4. Mitigating the risk of generative AI as a tool to undermine integrity

2.4.1. Generative AI can enhance the work of integrity actors, but it also creates the need for greater vigilance of evolving integrity risks

Generative AI poses unique risks to the work of anti-corruption and integrity bodies specifically. For instance, one category of risks is adversarial attacks. Broadly, adversarial attacks represent a cybersecurity vulnerability where attackers design inputs to evade detection. Generative AI can be used to create advanced phishing communications or enable actors with malicious intent to convincingly mimic individuals or entities, thus heightening the risk of identity theft, fraud and social engineering. Moreover, the spread of deepfakes (highly realistic videos, audios, or images) amplifies this threat (Shabsigh and Boukherouaa, 2023[2]).

Other risks range from LLMs making it easier for public officials to commit fraud to making it more difficult for integrity actors to detect corruption (Independent Commission Against Corruption, 2023[32]). The politically sensitive nature of anti-corruption work also means that leaning too heavily on automated decision-making can lead to the undermining public trust or augmenting ethical concerns. As it incorporates AI broadly and generative AI specifically into its work, the Independent Commission Against Corruption (ICAC) of New South Wales in Australia has considered how these and other threats could manifest. For more detail see Box 2.9.

Box 2.9. Insights from the Independent Commission Against Corruption (ICAC) of New South Wales on AI's potential threats to anti-corruption work

ICAC is an anti-corruption agency (ACA) in the Australian federal state of New South Wales that is responsible for a number of anti-corruption activities, including both the promotion of corruption prevention measures and investigation of corruption allegations. In response to a legislative inquiry on the use of AI in New South Wales in 2023, ICAC produced a report outlining both the opportunities and threats that AI poses to its work. Potential opportunities included the ability of AI to enhance intelligence through filtering, sorting and analysing large data sets; pattern recognition; forecasting and modelling; sentiment analysis; detection of anomalies in data; and data integration and multi-source analysis. ICAC also noted the ability of AI to reduce opportunities for corruption by limiting the degree of human discretion in decision making.

However, ICAC also noted that AI risks frustrating its anti-corruption efforts in several ways. These include:

1. The ability of AI to produce deepfakes
2. AI enhanced cybercrime
3. Exploitation of AI by public officials
4. Deference to AI
5. The use of AI to forge government documents
6. Threats to democracy and public discourse
7. Risks related to outsourcing

Points 3 and 5 are particularly noteworthy for the work of integrity actors in government, especially given the role of LLMs. First, ICAC notes that it has already investigated public officials who have tampered with IT systems to cover up corrupt conduct. LLMs could exacerbate this problem. An individual with enough technical expertise could poison data or manipulate models in order to alter system outputs.

They could also take advantage of known system vulnerabilities for personal gain or even sell information on these vulnerabilities to other corrupt actors.

ICAC also notes that many of its investigations relate to fraudulent documents, including procurement information, recruitment information, grant applications, building certificates, applications for business licenses, and conflict-of-interest declarations. While forging or altering these documents is difficult for humans to do in a convincing manner, it would be relatively easy for generative AI. There is also the risk that generative AI produces fraudulent documents without being prompted, resulting in fraud investigations against individuals with no malintent. In cases where fraud was premeditated, it is nonetheless easy for individuals to leverage the "black box" nature of advanced technology to feign ignorance.

These are just some of the threats that generative AI may pose to anti-corruption and integrity work. More broadly, ICAC notes that AI systems can reduce public trust in government decision making or may detach decision makers from those affected by their decisions to such an extent that they no longer consider moral ramifications. It is important that integrity actors keep these threats in mind as generative AI becomes more widely used and understood.

Source: (Independent Commission Against Corruption, 2023[32])

Going beyond LLMs, the OECD has previously raised issues concerning the risk that generative AI can amplify misinformation (i.e. the unintended spread of false information) and the deliberate spread of disinformation by malicious actors (Lorenz, Perset and Berryhill, 2023[3]). For instance, the widespread dissemination of false information during the COVID-19 pandemic highlighted the severe consequences disinformation can have on the execution of policies, as well as on trust and unity within society (Matasick, Alfonsi and Bellantoni, 2020[33]).

Generative AI also poses risks in the context of lobbying. As illustrated by previous examples, generative AI, particularly LLMs, can help entities to quickly process and provide inputs on draft legislation. Yet, generative AI as a form of AI also has the potential to completely overwhelm government consultation platforms by spamming them with fake or repetitive comments in order to amplify certain processes or stymy the policymaking process altogether (Smith and Harris, 2023[34]). The solutions are beyond the mandate of many integrity actors, but this issue is worth bearing in mind as they advance with relevant AI initiatives. Countries have done little to amend lobbying legislation to account for this threat, and opportunities remain to adjust the scope of lobbying legislation and the information that lobbyists are required to disclose in light of it. Countries can also invest in the necessary IT tools to manage increased comment volume and distinguish between legitimate and AI-generated comments.

While LLMs can amplify or create new risks that have implications for the work of integrity actors and their external environment, other risks are more internal in nature and can be considered when developing control and risk mitigation measures. Bias in AI can occur unintentionally, yet deliberate attempts to undermine or exploit LLMs for nefarious purposes present an evolving challenge. The OECD has reported previously on these issues in different contexts. For instance, based on the framework of the Surveillance Commission of the Financial Sector (*Commission de Surveillance du Secteur Financier*, CSSF) in Luxembourg, the OECD highlighted risk related to data poisoning and model theft, among others (Berryhill et al., 2019[20]).

- Data poisoning. This includes tampering with the training data, leading the AI to learn incorrect patterns. This is particularly problematic for types of AI that rely on continuously updated online data sources. For instance, individuals might create misleading content on social media to disrupt an AI's ability to accurately perform sentiment analysis. Similarly, subtle alterations to images, indiscernible to humans, can trick an AI into misidentifying new images.

- Model theft. This refers to the risk of unauthorised replication of a model, whereby attackers reverse-engineer an LLM or breach security measures to access proprietary or sensitive information. Examples include the hijacking of AI-powered chatbots for public services to create misleading or fraudulent services, or stealing of models used for predictive policing in order to circumvent law enforcement strategies.

While this paper concentrates on how integrity actors use generative AI, particularly LLMs in their operations, it is crucial to acknowledge that these technological advancements must be accompanied by a deeper awareness of the potential integrity risks they pose. The very risk assessments that can be enhanced by generative AI may, in a variety of government spheres, need to account for how the same technology can exacerbate integrity risks.

References

AI Sweden (2024), *A common digital assistant for the public sector*. [10]

Berryhill, J. et al. (2019), "Hello, World: Artificial intelligence and its use in the public sector", *OECD Working Papers on Public Governance*, No. 36, OECD Publishing, Paris, https://doi.org/10.1787/726fd39d-en. [20]

Bumann, J. and M. Peter (2019), *Action Fields of Digital Transformation - A Review and Comparative Analysis of Digital Transformation Maturity Models and Frameworks*, Edition Gesowip, https://www.researchgate.net/publication/337167323_Action_Fields_of_Digital_Transformation_-_A_Review_and_Comparative_Analysis_of_Digital_Transformation_Maturity_Models_and_Frameworks. [13]

Danish Agency for Digitalisation (2024), *Guide for public authorities on the responsible use of generative artificial intelligence*, Danish Agency for Digitalisation, Copenhagen. [27]

Emett, S. (2023), *Leveraging ChatGPT for Enhancing the Internal Audit Process – A Real-World Example from a Large Multinational Company*, https://ssrn.com/abstract=4514238 or http://dx.doi.org/10.2139/ssrn.4514238. [12]

European Parliament (2024), "Artificial Intelligence Act: MEPs adopt landmark law", *European Parliament News*, Press Release on 13 March 2024, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (accessed on 18 March 2024). [15]

Federal Council of Switzerland (2020), *Guidelines on Artificial Intelligence for the Confederation*, Federal Council of Switzerland, Bern. [26]

Gao, Y. et al. (2023), "Retrieval-augmented generation for large language models: A survey.", *arXiv preprint*, https://arxiv.org/abs/2312.10997. [17]

Gesnouin, J. et al. (2024), *LLaMandement: Large Language Models for Summarization of French Legislative Proposals*, https://arxiv.org/abs/2401.16182. [18]

Government of the Netherlands (2024), *The government-wide vision on Generative AI of the Netherlands*, Government of the Netherlands, The Hague. [6]

Huang, A. and H. Yi Yang (2023), "FinBERT: A Large Language Model for Extracting Information from Financial Text", *Contemporary Accounting Research*, Vol. 40/2, https://doi.org/10.1111/1911-3846.12832. [11]

Independent Commission Against Corruption (2023), *SUBMISSION TO THE LEGISLATIVE COUNCIL INQUIRY INTO ARTIFICIAL INTELLIGENCE IN NEW SOUTH WALES*, Independent Commission Against Corruption (ICAC), Sydney. [32]

Izdebski, K., T. Turashvili and H. Harutyunyan (2023), *The Digitalization of Democracy: How Technology is Changing Government Accountability*, National Endowment for Democracy, Washington. [24]

Köbis, N., C. Starke and I. Rahwan (2021), *Artificial Intelligence as an Anti-Corruption Tool (AI-ACT): Potentials and Pitfalls for Top-down and Bottom-up Approaches*, Max-Planck-Institute for Human Development, Center for Humans and Machines. [21]

Liao, V. and J. Vaughan (2023), *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap*, https://arxiv.org/pdf/2306.01941.pdf. [31]

Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence"*, OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en. [3]

Matasick, C., C. Alfonsi and A. Bellantoni (2020), "Governance responses to disinformation: How open government principles can inform policy options"*, OECD Working Papers on Public Governance*, No. 39, OECD Publishing, Paris, https://doi.org/10.1787/d6237c85-en. [33]

Odilla, F. (2023), "Bots against corruption: Exploring the benefits and limitations of AI-based anti-corruption technology", *Crime, Law and Social Change*, Vol. 80/4, pp. 353-396, https://doi.org/10.1007/s10611-023-10091-0. [22]

OECD (2023), "AI language models: Technological, socio-economic and policy considerations"*, OECD Digital Economy Papers*, No. 352, OECD Publishing, Paris, https://doi.org/10.1787/13d38f92-en. [1]

OECD (2023), *Global Trends in Government Innovation 2023*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/0655b570-en. [23]

OECD (2023), "Recommendation of the Council on Artificial Intelligence"*, OECD Legal Instruments*, OECD/LEGAL/0449, OECD, Paris, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [25]

OECD (2022), "Facilitating citizen and stakeholder participation through the protection of civic freedoms", in *The Protection and Promotion of Civic Space: Strengthening Alignment with International Standards and Guidance*, OECD Publishing, Paris, https://doi.org/10.1787/9ca8987d-en. [19]

OECD (2022), *Strengthening Analytics in Mexico's Supreme Audit Institution: Considerations and Priorities for Assessing Integrity Risks*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/d4f685b7-en. [4]

OECD (2021), *Countering Public Grant Fraud in Spain: Machine Learning for Assessing Risks and Targeting Control Activities*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/0ea22484-en. [8]

OECD (2020), *Good Practice Principles for Data Ethics in the Public Sector*, OECD, Paris, https://www.oecd.org/gov/digital-government/good-practice-principles-for-data-ethics-in-the-public-sector.pdf. [14]

OECD (2020), "The OECD Digital Government Policy Framework: Six dimensions of a Digital Government", *OECD Public Governance Policy Papers*, No. 02, OECD Publishing, Paris, https://doi.org/10.1787/f64fed2a-en. [35]

OECD (2019), "Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449", *OECD Legal Instruments*. [28]

OECD (2014), "Recommendation of the Council on Digital Government Strategies", *OECD Legal Instruments*, OECD/LEGAL/0406, OECD, Paris, https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf. [37]

Office of the Auditor General of Norway (2018), *Auditing to benefit the society of tomorrow: Strategic plan 2018–2024*, Office of the Auditor General of Norway, Oslo. [5]

Otia, J. and E. Bracci (2022), "Digital transformation and the public sector auditing: The SAI's perspective", *Financial Accountability & Management*, Vol. 38/2, pp. 252-280, https://doi.org/10.1111/faam.12317. [36]

Shabsigh, G. and E. Boukherouaa (2023), "Generative Artificial Intelligence in Finance: Risk Considerations", *Fintech Notes*, Vol. 2023/006, https://doi.org/10.5089/9798400251092.063. [2]

Smith, A. and M. Harris (2023), *How artificial intelligence and large language models may impact transparency*, Westminster Foundation for Democracy. [34]

U.S. Government Accountability Office (2024), *Artificial Intelligence Use Cases*, https://www.gao.gov/science-technology/artificial-intelligence-use-cases. [9]

UK Government (2024), *Guidance: Generative AI Framework for His Majesty's Government*, https://www.gov.uk/government/publications/generative-ai-framework-for-hmg/generative-ai-framework-for-hmg-html. [16]

US Government Accountability Office (2021), *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, US Government Accountability Office, Washington. [29]

World Bank (2023), *The Governance Risk Assessment System(GRAS) Advanced Data Analytics for Detecting Fraud, Corruption, and Collusion in Public Expenditures*, https://openknowledge.worldbank.org/handle/10986/40640. [7]

Zhao, H. (2023), "Explainability for Large Language Models: A Survey", *ACM Transactions on Intelligent Systems and Technology*, https://doi.org/10.1145/3639372. [30]

Notes

[1] This designation is without prejudice to positions on status and is in line with United Nations Security Council Resolution 1244/99 and the Advisory Opinion of the International Court of Justice on Kosovo's declaration of independence.

[2] We estimate that over 150 organisations received the questionnaire, but we did not attempt to track the total number of recipients given the qualitative purpose of our research. The aim of our questionnaire was to collect insights and use cases from a targeted group of government entities (i.e. integrity actors) rather than achieving statistical representativeness. Without additional information as to why some recipients decided not to complete the questionnaire, reporting on the total number of recipients does not materially contribute to the qualitative nature of our findings.

[3] LangChain is a framework designed to build applications powered by language models, facilitating the creation of context-aware applications that connect to various sources for context—such as prompt instructions, examples, and content grounding—and utilise language models for reasoning, including determining responses based on context and deciding on actions to take. See https://www.langchain.com/.

[4] For instance, see: OECD (OECD, 2023[25]), "Recommendation of the Council on Artificial Intelligence" https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; OECD (2014[37]), "Recommendation of the Council on Digital Government Strategies", https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf; OECD (2020[35]), The OECD Digital Government Policy Framework: Six dimensions of a Digital Government; OECD (2022[4]), *Strengthening Analytics in Mexico's Supreme Audit Institution: Considerations and Priorities for Assessing Integrity Risks*; OECD (2021[8]), *Countering Public Grant Fraud in Spain: Machine Learning for Assessing Risks and Targeting Control Activities*; Otia and Bracci (2022[36]), Digital transformation and the public sector auditing: The SAI's perspective; and Bumann and Peter (2019[13]), *Action Fields of Digital Transformation - A Review and Comparative Analysis of Digital Transformation Maturity Models and Frameworks*.

[5] Software, Tools, and Compliance, as a response option, covered essential software, development tools, and adherence to legal/regulatory standards. System Scalability and Integration pertains to the ability to scale IT resources and integrate the LLM with existing technology stacks.

[6] See, for instance, XAITK, an open-source explainable AI toolkit built with the support of the Defense Advanced Research Projects Agency (https://xaitk.org/ and https://www.darpa.mil/program/explainable-artificial-intelligence).