



Economic Commission for Europe**Conference of European Statisticians****Seventy-second plenary session**

Geneva, 20 and 21 June 2024

Item 5 of the provisional agenda

**Use of Artificial Intelligence and Large Language Models
in official statistics and authoritative geospatial data****Working together to advance artificial intelligence for official
statistics: an overview of the High-Level Group for the
Modernisation of Official Statistics initiatives and outputs****Note by the Large Language Models Team under the High-Level
Group for the Modernisation of Official Statistics and the Secretariat***Summary*

This document provides an overview of the initiatives on artificial intelligence (AI) undertaken by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) from 2019 to the present, and outputs derived from collaborative efforts. The annex is an excerpt from the white paper “Large Language Models for Official Statistics” prepared by a dedicated task team under the High-Level Group for the Modernisation of Official Statistics. The full white paper is available [here](#).

The document is presented to the Conference of European Statisticians’ session on “Use of artificial intelligence and large language models in official statistics and authoritative geospatial data” for discussion.



I. Introduction

1. Artificial intelligence (AI) holds a great potential for statistical organizations. It can make the production of statistics more efficient by automating certain processes or assisting humans to carry out the processes. Additionally, AI enables statistical organizations to leverage new types of data such as social media data and imagery, allowing them to provide society and policymakers with more detailed and timely insights.
2. Many statistical organizations have begun adopting this new technology to enhance the relevance and quality of official statistics. However, as it often happens with new technology in its early stage, each organization finds itself with only limited resources to navigate the full potential of the technology on its own. This makes pooling experiences and knowledge from different organizations invaluable to facilitate the adoption of the new technology.
3. The High-Level Group for the Modernisation of Official Statistics (HLG-MOS) has served as an international platform for practitioners in national and international statistical organizations to exchange experiences and lessons learned, co-create best practices, and develop frameworks to navigate this fast-evolving landscape together. Initially, its focus was on machine learning (ML), a subset of AI which involves training computer algorithms to learn from data and make predictions or decisions without explicit programming, but it has since expanded its scope to include a more recent developments in the field, such as large language models (LLMs).
4. This note aims to provide an overview of the initiatives undertaken by HLG-MOS from 2019 to the present, and outputs derived from collaborative efforts.

II. Work on machine learning (2019–2022)

5. The discussion on ML began to emerge within the HLG-MOS community in the mid-2010s, for example, in the context of big data. However, it gained significant momentum following the position paper on ML by the Blue-Skies Thinking Network (BSTN), the HLG-MOS Modernisation Group on horizon scanning, in 2018. The paper led to an HLG-MOS Project on ML, which subsequently led to the establishment of the Machine Learning Group spearheaded by the Office for National Statistics (ONS) of the United Kingdom of Great Britain and Northern Ireland.

- **HLG-MOS Machine Learning Project (2019–2020)**: the project was launched under HLG-MOS in March 2019 and concluded at the end of 2022. The project was structured around three work packages: (1) pilot studies; (2) quality; and (3) integration. The project started with a small group of 11 participants but gradually grew into a large community of interest with over 120 participants from 37 national and international organizations. More details about the project can be found on the [ML Project wiki page](#).
- **ONS-UNECE Machine Learning Group (2021–2022)**: the large community of interest built from the ML Project could continue thanks to ONS who volunteered to coordinate further international collaboration on ML. Compared to the ML Project, which focused more on experimentation around ML (“what can we do with ML?”), the integration issue (“how can ML be integrated into regular work?”) was given more emphasis during ML Group’s work. In addition to continuing pilot studies with more diverse use cases (e.g., web-scraping data, Automatic Identification System data, small area estimation, route optimization), the ML Group covered topics such as challenges around moving from experimentation to production, ethical considerations, quality of training data and infrastructure. The ML Group brought together more than 400 people from 35+ different countries and 20+ international organizations. More details about the work of the ML Group in each year can be found in the [ML Group 2021 wiki page](#) and the [ML Group 2022 wiki page](#).

6. A lot of outputs have been produced over these 4 years. For the ease of navigating the extensive pool of resources, they are structured according to typical questions raised about ML in the rest of the section.

A. What is machine learning?

7. Use of machine learning in the context of official statistics is still relatively new. Training on what machine learning is and in relevant skills is paramount for the adoption of this technology in statistical organizations. Many training resources related to ML exist outside of official statistics community. A few introductory resources prepared under the umbrella of HLG-MOS include:

- UNECE publication [“Machine Learning for Official Statistics” \(2022\) – Chapter 2](#)
- Introductory learning resources recommended by ML project participants which were compiled on the [“Learning and Training” wiki page](#).

B. Where machine learning can be used in statistical organizations?

8. Machine learning can contribute to the work of statistical organizations in various ways. It can automate processes which used to be largely done by humans and allow the organizations to make use of new data sources. The utilization of big data often requires ML as it involves processing a vast amount of data efficiently and effectively. Main areas of applications explored include:

- Coding and classification: [ML Project Classification and Coding Theme Report \(2020\)](#) and [ML Group Text Classification Theme Group Report \(2022\)](#); [ML Group Web Scraping Theme Group Report \(2022\)](#)
- Edit and imputation: [ML Project Edit and imputation Theme Report \(2020\)](#)
- Image analysis: [ML Project Imagery Theme Report \(2020\)](#).

9. Studies and accompanying codes (where available) under these application areas from 2018–2022 are compiled in the [“Studies and Codes” wiki page](#) along with other use case examples. Session 1 of the [UNECE Workshop on ML \(2023\)](#) also presents more recent examples of ML applications. The [“knowledge transfer” report \(2021\)](#) touches on enablers and blockers of replicating ML examples across different statistical organizations.

C. What are the quality implications and ethical considerations needed?

10. Ensuring quality is non-negotiable for producers of official statistics. However, the so-called “black box” nature of machine learning and its heavy dependence on training data makes it more difficult for statistical organizations to guard against bias and errors. Therefore, quality and ethical implication was one of most significant concerns from the beginning of ML exploration. Following outputs explored quality and ethical considerations:

- Quality framework for statistical algorithm ([UNECE publication “Machine Learning for Official Statistics” \(2022\) – Chapter 4](#)) and [its application \(2021\)](#)
- [Quality of training data \(2022\)](#)
- [Ethical Consideration in the Use of ML for Research and Statistics \(2021\)](#).

D. What are the organizational challenges in moving from experiments to production?

11. Despite successful pilot studies, integrating ML models into production often proves difficult and time-consuming, with many solutions from experiments left on the shelf. To ensure successful deployment, proactive planning is essential, addressing organizational and cultural challenges. Key outputs exploring these challenges include:

- [Journey from Experiment to Production \(2021\)](#)
- [Organizational aspects of implementing ML-based data editing in statistical production \(2024\)](#).¹

E. How machine learning capability can be integrated in statistical organizations?

12. The ability to use and scale ML requires more than just the underlying ML code. After initial experimentation and proof-of-concept, it is important to establish a structure that allows to expand the use of ML beyond a small group of experts. One key aspect of building sustainable capacity for ML is the focus on MLOps – the practice of integrating ML models into the production environment in a systematic and efficient manner. This involves building the infrastructure and platforms necessary to deploy, monitor, and maintain ML models at scale. Following outputs explored these issues:

- Model retraining reports [from 2021](#) and [from 2022](#)
- [Building an ML Ecosystem in Statistical Organizations \(2022\)](#).

III. Work on large language models (2023)

13. The capabilities of AI have made a significant leap forward in the last few years with the advance of large language models (LLMs) and there is a growing recognition of the transformative potential of LLMs in the statistical community.

14. In mid-2023, two HLG-MOS Modernisation Groups – Blue-Skies Thinking Network (BSTN) and Applying Data Science and Modern Methods (ADSaMM) Group drafted a white paper “LLM for Official Statistics”.

15. The paper provides a short introduction to LLMs, an overview of where they can be used in the statistical production, and associated risks, concrete uses cases from five statistical organizations, main considerations that should be taken into accounts for moving forward with this new technology. The full paper is available [here](#) and an excerpt from the paper is provided in the Annex.

IV. Work in progress (2024–)

16. In 2024, HLG-MOS started several streams of work aimed at integrating AI applications into official statistics.

A. High-Level Group for the Modernisation of Official Statistics project on Generative AI for Official Statistics

17. Building on the LLM white paper, the project aims to further investigate the potential of generative AI, a broader category of advanced AI system that encompass LLMs. Along with concrete uses cases of generative AI in statistical organizations (e.g., retrieval-augmented generation, code generation), the project will explore the following aspects: project management and development journey (organizational aspects), prompt engineering (best practices and quality assurance), architecture and applications stack (infrastructure, tools) and governance and ethics (analysis of security, legal, and ethical risks and mitigation).

¹ The output is from the HLG-MOS Applying Data Science and Modern Methods Group, but it is included as it is highly relevant. See Section IV for more information about the Group.

B. Applying Data Science and Modern Methods Group

18. The Applying Data Science and Modern Methods (ADSaMM) Group was launched in early 2022, reflecting the increasing importance of new data sources and methods for the compilation of official statistics. The Group aims to go beyond conceptual frameworks for data science and modern methods, identifying concrete opportunities to further modernize business processes of statistical organizations. In 2024, following topics will be explored under ADSaMM Group:

- Advancement of responsible AI in statistical offices: building on the “Framework for Responsible AI/ML for Official Statistics”² from 2023, the task team will work on promoting ethical AI deployment in statistical practices, aligning with fairness and transparency principles. The work will include developing guidelines and toolkits for ethical AI application as well as enhancing understanding and implementation of responsible AI among statistical organizations through training and workshops.
- Uncertainty Quantification: ML-based results are often presented without measures for uncertainty which raises concerns about the reliability. Rigorous ways of qualifying the uncertainty can help statistical organizations to provide increased assurance when communicating the use of ML/AI to both users and the general public. The task team will conduct research on traditional approaches (e.g., Bayesian, bootstrap) and conformal prediction approaches.

C. Capability and Communication Group

19. The Capability and Communication (CapComm) Group is an HLG-MOS Modernisation Group that focuses on the organizational changes and the communication challenges necessary to support modernization in statistical organizations. The Group created a task team “use of AI for official statistics from communication perspective” to explore how to enhance productivity of communication experts with AI among others and how to communicate the use of AI to maintain trust in official statistics.

² Under finalisation.

Annex

White paper “Large Language Models for Official Statistics” by the High-Level Group for the Modernisation of Official Statistics

Note that this Annex includes only Section 1 (“Introduction on Large Language Models”), Section 2 (“Implication and opportunities for official statistics”) and Section 5 (“Considerations as statistical organizations move forward with Large Language Models”) from the LLM paper due to the word limit constraints of the current document. For the remaining sections – Section 3 (“Use cases for statistical organizations”) and Section 4 (“Risks and mitigation measures”) – please refer to the [full paper](#).

Executive summary

Large language models (LLMs) are a class of artificial intelligence that can understand, interpret and generate texts. Based on the extensive training on vast data sets with billions of parameters, LLMs are capable of understanding and generating texts at a level indistinguishable from humans. This sets them apart from traditional machine learning models whose application is primarily focused on assisting humans in prediction tasks rather than creating content.

There is little doubt that LLMs are going to play an important part in statistical organizations’ operations into the future. Like any offices in many sectors and domains, statistical organizations have regular workplace tasks such as writing emails and meeting notes. LLMs could assist staff with these routine but time-consuming duties. Moreover, LLMs can be used to enhance efficiencies at various stages of statistical production processes and other related work, provided with human supervision and careful examination against existing methods. These opportunities are not just theoretical, but very much real. Implementation examples from various national and international organizations on use cases such as SAS to R translation, statistical classification system updates, report generation, natural languages-based data search and editing of metadata demonstrate this.

However, there are risks arising with LLMs such as ethical issues, legal implications (such as copyright) and a general lack of awareness and literacy. Also, due to its very capability to generate texts that are very well-written and contextually relevant, users could be misled to factually incorrect, outdated and even entirely fabricated (called “hallucination”) data. Privacy and security concerns regarding potential data leaks through LLMs are of a great concern for statistical organizations as well. These risks are often dependent on the types of use cases LLMs are employed for, but there are general mitigation measures such as ensuring human oversight, using language testing protocols, local fine-tuning and application of privacy principles and requirements.

As statistical organizations move forward, there are several main considerations that should be taken into account. These include how to establish a governance structure, how to engage with tech companies that provide the LLMs, services based on LLMs and cloud computing, as well as how to select LLMs with varying levels of openness. Given the heightened public interest and scrutiny faced by public organizations, communicating the responsible use of LLMs – that statistical organizations are using them purposely where there are clear benefits with awareness of risks and necessary mitigation measures – is vital. The use of LLMs by statistical organizations is still in its infancy, but there are a few practical suggestions:

- provide training on LLMs at all levels in the organization (technical, operational, and managerial)
- approach LLMs with the execution of small pilot projects to gain familiarity with the technology and understand the potential value
- develop an overarching LLM strategy once awareness and familiarity have reached a sufficient level, and

- devote continuous effort to keep up to date with the continuously changing landscape of LLMs.

Due to the dynamic and fast-evolving nature of this field, a close collaboration among statistical organizations will continue to be crucial to collectively explore different applications and share insights and experiences along their journey.

1. Introduction on large language models

Large language models (LLMs) are still a relatively new technology. Therefore, it is important to understand what they are and how they work before delving into the implication of LLMs for official statistics. The focus of this section is to explain the capabilities of LLMs, their roots in the broader artificial intelligence landscape, and their transformative power in natural language processing (NLP). We will briefly describe the dynamic evolution of language models, from the complexity of transformer neural networks to the adaptability of basic models such as Bidirectional Encoder Representations from Transformer (BERT) and Generative Pre-trained Transformer (GPT). We will then briefly discuss the concepts that are important for LLMs such as fine-tuning models and prompt tuning that improves the capabilities of LLMs without having to retrain them from scratch, and open source in LLMs.

1.1. What are large language models?

LLMs are a class of artificial intelligence (AI) that can understand, interpret, and generate texts. Based on the extensive training on vast data sets, LLMs are capable of understanding and generating texts at a level indistinguishable from humans. LLMs have become increasingly popular due to their exceptional ability in a wide range of NLP tasks and natural language understanding (NLU) tasks, such as translation and text summarization.

In services developed based on LLM (e.g., ChatGPT), users can interact with LLMs through natural languages, called “prompts” (instruction that generates responses from LLMs), for example, as below:

User: Could you give me excel functions that generate random integer numbers between 1–10?

LLM service: Certainly! You can use RANDBETWEEN function. RANDBETWEEN(1, 10) generates a random whole number between the specified minimum and maximum values.

User: How about if I want a real number between 0 and 10?

LLM service: If you want a real number (including decimals) between 0 and 10, you can use the RAND() function and then scale the result.

Relationship with artificial intelligence, machine learning and generative artificial intelligence

LLM is not a sudden, new technology that emerged out of nowhere; it is the culmination of the continuous development and evolution of AI. To better understand the essence of LLMs, it is important to comprehend the context of their creation and the differences between various technologies and definitions. Artificial intelligence, machine learning, large language models, and generative AI are all interconnected concepts, but there are crucial distinctions among them. Before focusing on LLMs, we will examine the closely connected concepts.³

- **Artificial intelligence (AI)** is a broad field of computer science that focuses on creating systems and machines capable of performing tasks that typically require human intelligence. These tasks include problem-solving, learning, reasoning, perception, language understanding, and more.
- **Machine learning (ML)** is a subset of AI that involves the use of algorithms and statistical models to enable computers to improve their performance on a specific task

³ <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8259629>.

through learning from data, without being explicitly programmed. In other words, it's about teaching computers to learn from examples and make predictions or decisions based on that learning. Many AI applications use ML techniques to achieve their goals.

- **Deep learning** is a subset of ML that employs artificial neural networks with many interconnected layers (deep neural networks). These networks can automatically discover and learn to represent patterns or features from large volumes of data. Deep learning has been highly successful in tasks like image and speech recognition. It is particularly well suited for tasks involving complex, unstructured data like images, audio, and text. It is a specialized tool within the ML toolkit.
- **Generative AI** refers to AI systems that can generate new content or data that is not explicitly derived from existing examples. This can include generating text, images, music, and more. Generative AI often uses techniques like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

LLMs such as GPT-3 are a specific application of deep learning within the field of AI. They are capable of natural language understanding (i.e., they use algorithms and models capable of accurately interpreting human language) and generation and are used in a wide range of applications. Modern LLMs emerged in 2017 and use transformer neural networks, commonly referred to as transformers. With a large number of parameters and the transformer model, LLMs are able to understand and generate accurate responses rapidly, which makes the AI technology broadly applicable across many different domains.

LLMs can be viewed as a subset of foundation models⁴ that focus on language-related tasks. Foundation models are large deep learning neural networks trained on large datasets and serve as a fundamental building block for various applications. They can produce wide and various outputs (e.g., text, image and audio) or can generally apply a pretraining objective to the dataset, so that the foundation model becomes good at that objective (e.g., image creation). Foundation models can be used as a 'base' for other models, which can be built on top of the foundation model. A foundation model is so large and impactful that it serves as the foundation for further optimization and specific use cases.

1.2. How do large language models work?

Large language models are based on deep learning architectures, with a specific focus on so-called transformer models. Transformers are neural network architectures that use self-attention mechanisms to process input data, enabling them to handle long-range dependencies in language effectively. The following sections detail the components and training process of large language models.

Components of large language models

- (1) **Parameters:** The core components of a large language model are its parameters, which include weights and biases. These parameters are adjusted during training to minimize the difference between the model's predictions and the real values.
- (2) **Layers:** A large language model comprises several layers, each responsible for extracting and processing different levels of information from the input data. These layers typically include input and output layers, as well as multiple hidden layers.
- (3) **Attention mechanism:** The attention mechanism is a critical component of large language models, allowing them to selectively focus on relevant parts of the

⁴ Foundation model is defined as "any model that learns from rich data (typically using self-supervised learning) that can be adapted (e.g., fine-tuned) for a wide range of downstream applications" by the Center for Research in Fundamental Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI) <https://hai.stanford.edu/news/reflections-foundation-models>. What makes foundation models unique is their general nature and size, which set them apart from traditional machine learning models. They can be used as a basis for the development of specialized subsequent applications.

input data. This mechanism helps the model capture dependencies between words and phrases, even when they are far apart in the text.

Large language models are trained on a massive dataset, usually containing billions of words from diverse sources. This self-supervised learning process enables the model to learn the structure and patterns of the language. Training a large LLM such as GPT-3, which has 175 billion parameters, is a very expensive process that can cost tens of millions of dollars in hardware and electricity costs alone. However, the pre-trained models can be fine-tuned on a smaller, task-specific dataset. This “fine-tuning” process refines the model’s understanding of the specific task, helping it generalize better and achieve higher performance on that task. The fine-tuning step still requires sufficient computing power for the given model and task but is less resource intensive compared to pre-training the models from scratch. The fine-tuning is further described in the following section.

1.3. Fine-tuning models

LLMs are often used *off the shelf* (i.e., they come pre-trained with a full set of weights). It can nevertheless be possible to customize them, by using a number of techniques including Prompt tuning and fine-tuning, which can improve model output without the need to fully train them from scratch.

Prompt tuning is a “lightweight” method that involves designing specific inputs to guide the model’s output. This tuning is done without changing the model’s parameters. Prompt tuning capitalizes on the model’s existing knowledge and capabilities by simply altering the way it’s queried.

Fine-tuning is a more intensive process that entails training the model, after the initial training, on a specialized dataset. This is aimed at improving the model’s performances for the target task or domain. During fine-tuning, the model’s parameters are updated to better align with the target task or domain. This allows the model to produce more relevant and specific outputs for specialized applications. However, fine-tuning requires additional computational resources and a curated dataset for effective results.

Fine-tuning models can be an effective approach for:

- Refining style and expression: Tailor the model’s output to match specific styles, tones, formats, or other desired qualitative aspects (e.g., a statistical chatbot).
- Addressing complex instructions: Effectively handle intricate and detailed prompts, even those that involve multiple steps.
- Boosting consistency: Improve the model’s reliability in producing the intended outputs, reducing the likelihood of unexpected or undesirable results.
- Mastering new skills: Enable the model to perform new tasks or skills that may be challenging to define explicitly in a prompt format.

From the technical point of view, the dataset needed to fine-tune a model is often structured as a JavaScript Object Notation (JSON) file. It is organized in triplets of roles (defined as “system”, “user” and “assistant”) each one associated with a content. The triplets define the model behaviour in a given situation; each triplet system/user/assistant is used to recalibrate the model.

In the example below, a chunk of the dataset used for fine-tuning a chatbot using ChatGPT 3.5 is shown; three triplets are presented (for brevity, the “System” prompt is shown only once, being the same for the three cases).

```
{"messages": [{"role": "system", "content": "You are a support agent whose primary goal is to help users. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to official statistics."}, {"role": "user", "content": "What is the best country in Australia?"}, {"role": "assistant", "content": "I apologise, but I cannot provide personal opinions. My primary job is to assist you with queries related to official statistics. Do you have any official statistics topic that need addressing?"}]}
```

```
{"role": "user", "content": "Where I can find statistics on National Accounts?"}, {"role": "assistant", "content": "To find statistics on National Accounts you can connect to the Data Browser and search for National account in the dropdown menu. Here the link: —./—"}]}
```

```
{"role": "user", "content": "I'm having trouble connecting to the Data Browser"}, {"role": "assistant", "content": "No worries, let's do a preliminary check. What browser are you using?"}]}
```

1.4. Open source

The term open source refers to something people can modify and share because its design is publicly accessible.⁵ “Open” in the open-source term is there for a reason. An open source LLM is a model whose code is made publicly available under an open-source licence, allowing anyone to use, adapt and share the model. Such open-source models are usually accompanied by detailed documentation that offers valuable information about the model structure, training methods, model configurations, and datasets used during training and evaluation. This documentation promotes a deeper understanding of the model’s internal workings and capabilities, promoting transparency and collaboration within the AI and ML community.

Thanks to this openness, users can, in addition to directly using the model, explore its design, adapt and customize the code, thus improving the model. This may be one of the opportunities for international collaboration between national statistical producers. A number of open-source models are available via Hugging Face – a company and a popular platform in the field of natural language processing (NLP) and artificial intelligence.

However, as discussed in detail in Section 5.3, users should carefully check the licence with which a LLM is made available to understand if their use case may be compliant with it. Several LLMs, for instance, were made publicly available with licences restricting commercial use. Other licences, instead, may impose users to publicly share derivative works with the same conditions of the original LLM, or require the user to explicitly credit the original creator. In summary, the fact that a LLM may be publicly available does not necessarily mean that there are no rules governing its use.

2. Implication and opportunities for official statistics

The possibilities for using LLMs are impressive but not unlimited, therefore it is important to understand what LLMs can and cannot do. In this section, we provide a general overview of how LLMs can improve the efficiency of routine tasks in statistical organizations, from communication to project management, highlighting their role in optimizing operations. The potential of LLMs to improve the efficiency of the statistical production process, from survey design to data dissemination will be discussed. We also give a closer look to the changing information landscape and how LLMs could affect the way people access statistical information.

2.1. What statistical organizations can do and cannot do with large language models

LLMs are trained on enormous amounts of information and contain billions of parameters to produce statistical predictions. Algorithms used in commercially available LLMs are rarely shared, leading to them being considered black boxes. As well, the training data is not clearly identified and could contain unintentional biases. Unfortunately, these biases could be reproduced in the results produced by an LLM. In addition, because the goal of the LLM is to predict the next word, they can produce incorrect or nonsensical information, commonly

⁵ <https://opensource.com/resources/what-open-source>.

referred to as hallucinations. However, since the outputs are very well written, human nature leads people to believe it as factual.

Despite these potential pitfalls, LLMs have many potential uses in statistical organizations. They are very good at understanding textual information, summarizing large amounts of information, and generating human-like responses which could be useful in automating many tasks within a statistical organization. This section will present some ideas where LLMs could be used, including tasks that are needed in any organization such as drafting emails and preliminary reports, summarizing information for brainstorming sessions, project management and translation to multiple languages. Also tasks that are particularly relevant for statistical organizations such as text classification, data visualization and data dissemination.

More details on these potential uses, and others, are presented later in the section.

While LLMs have the potential to change how statistical organizations work, **they must be closely monitored**. Emails and reports drafted by LLMs must be reviewed by humans to ensure that the context is correct and does not represent a biased viewpoint. This is important as LLMs are good at producing well written text, but they are not designed to verify that the content is factual or necessarily the best choice. If the data that the LLM is trained on is incorrect or only somewhat appropriate, it will use that information in formulating its response.

For example, the Applying Data Science and Modern Methods group of the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) posed several methodological questions to multiple LLMs and validated the results. In general, the responses were correct but not always the most appropriate. When asked about replacing missing values, a common response was to use mean imputation. While not incorrect, mean imputation is known to have some shortcomings such as distorting the distribution of the data and not using any auxiliary information that might be available. The questions posed by the group illustrated the fact that an LLM is a ‘reasoner’ and, unlike human experts, does not pose any questions to gather more information to find more suitable responses. The responsibility falls on the person querying the LLM to pose the correct prompts.

If the user is not knowledgeable in the subject, the LLM may not provide high quality responses. One of the essential tasks of a consultant is to work with a client to establish their real needs. In the context of a statistical consultant, this comes down to understanding the data needs and the ultimate use of the data to fill information gaps. This information is essential to ensure that the methods applied allow the data to fulfil the needs of the client. Without gathering this additional information, LLMs could suggest methods which may not be appropriate. If the individual interacting with the LLM has some subject matter knowledge, they will be able to provide additional information to arrive at an appropriate method. However, if the individual does not have the knowledge and they follow the advice of the LLM, the method put in place may not adequately solve the problem at hand.

This underlines the importance of **prompt engineering**, which requires some knowledge of the subject being discussed and understanding of how to get the best output from the LLM. In other words, LLMs will not be able to replace the human interaction required to clearly define the needs or the research question that is needed to arrive at the most appropriate statistical method. In the hands of a person who may not have a knowledge of the subject, blindly applying the advice of an LLM could lead to less than desirable results.

2.2. Improve efficiencies of regular workplace tasks

Like any other organization, statistical organizations have regular tasks that are quite similar to those found in both the public and private sectors. These tasks include activities such as managing emails, creating reports and presentations, and keeping meeting notes. Although these routine duties are vital for the organization to function effectively, they require a significant amount of time and effort from dedicated staff.

LLMs/ChatGPT can help streamline operations within statistical organizations and increase productivity of existing resources. This way the office can allocate its resources more

efficiently towards essential tasks and contribute to its objective of providing accurate and timely statistical information. In the following section, examples of how LLMs/ChatGPT can be employed to boost the efficiency of a statistical organization will be provided, enabling it to achieve its fundamental goals more effectively.

- (1) **Communications** – One of the most widespread applications of LLMs is the immediate application of its features into the communications. LLMs have proven to assist in drafting emails, plans, and reports by providing content suggestions, formatting help, and generating the text itself.⁶ This saves time and increases the quality of written materials. For reports, LLMs summarize lengthy documents, provide options for data visualizations, identify errors, and offer recommendations.
- (2) **Brainstorms and idea generation** – LLMs can facilitate brainstorming sessions by offering creative suggestions, exploring various angles of a problem, and generating new ideas based on the input provided. This can be particularly useful for diverse perspectives, exploring problem angles, prompting questions to deepen analysis, idea evaluation, shaping findings, and saving time.
- (3) **Project management and planning** – LLMs can be effectively used with routine tasks needed on various stages of the project management process by automating task planning and dependency management, optimizing resource allocation based on historical data and project requirements, estimating task durations for timeline planning, and simplifying meeting notetaking through transcription and summary generation, ensuring essential information is effectively documented and summarized.
- (4) **Translation from/to other languages** – LLMs can translate documents and text from one language to another, easing the access to information in different languages. Generally, at the current stage of development LLMs/ChatGPT offer significant advantages in translation tasks being more sensitive to the context. However, traditional automatic translation systems still hold advantages in scenarios involving large datasets, speed and efficiency requirements, and well-defined domains.
- (5) **Presentations** – LLMs can be used to create presentations from basic to advanced slides with macros. It can be employed not only for slides content generation, allowing customization of style, structure, and slide quantity, but also for developing talking points for a more human-friendly presentation tone.
- (6) **Educational purposes** – LLMs can be employed for educational and training purposes within the organization. They can provide explanations, create quizzes, and assist in designing e-learning materials to enhance the skills and knowledge of the workforce.
- (7) **Image generation** – Stock imagery is often used with reports and productions of statistical organizations. Rather than purchase stock imagery, statistics organizations could use LLMs to generate images to go along with statistical productions.

Adopting the wise use of LLMs/ChatGPT can free up human resources for more strategic and complex tasks, allowing staff to be more creative, productive, and put a focus on higher-priority areas.

⁶ Note that LLM-generated emails may be easy to recognize; friendly tip – do not copy paste generated text from the ChatGPT directly without style formatting as it will save the original font and grey background.

2.3. Improve efficiencies of statistical production and quality of service delivery

LLMs can be used in a wide array of applications to enhance efficiencies at various stages of statistical production process, provided with human supervision and careful examination against existing methods and expertise amassed in the organizations, for example,

- Design collection (GSBPM⁷ sub-process 2.3): LLMs can contribute to the design of surveys and questionnaires by suggesting questions, formats, and wording that are more likely to yield accurate responses.
- Classify and code (GSBPM sub-process 5.2): LLMs have the capability to automatically sort textual data into predefined categories or labels. Statistical organizations can use them for organizing survey responses and other textual data into pertinent categories in the statistical classification systems.
- Validate and edit data (GSBPM sub-process 5.3 and 5.4): LLMs can streamline data cleaning and pre-processing tasks by identifying and rectifying data errors, missing values, and inconsistencies.
- Produce dissemination products (GSBPM sub-process 7.2): LLMs can generate textual descriptions from a table or a series of numbers (see use case in Section 3.4. Report Generation Using LLMs (Statistics Canada)) which can be tailored to different audience segments, including policymakers, journalists, and the general public. This could greatly simplify the work of analysts and communication experts by providing initial drafts that human experts could work on. LLMs can also assist in automating the creation of charts and graphs, although this area is still under exploration.
- Metadata plays a crucial role in statistical production and editing of metadata can be assisted by LLMs (see use case in Section 3.5. Metadata Editing Leveraging GPT (Bank of International Settlements)).

In addition to their applications in the statistical production process, LLMs can provide support in several cross-cutting areas that are crucial for statistical organizations:

- Assist coding and translating between programming languages: LLMs can deal with not only natural languages but also programming languages which statistical organizations extensively use for many parts of its production, in particular, for processing and analysis. LLMs could significantly enhance the efficiency and effectiveness of programmers and analysts by helping streamlining and optimizing code development, providing code snippets and translating between different programming languages (see use case in Section 3.2. Code Translation and Explanation (SAS to R) Using LLMs (Ireland Central Statistics Office)).
- Update and maintain statistical standards: generate draft text descriptions to assist human experts in updating statistical classification systems (see use case in Section 3.1. Updating Statistical Classification Definitions (Australian Bureau of Statistics)) and methodology documents.
- Generation of synthetic data: Privacy and data use are key concerns when testing statistical methodology. LLMs can be used to generate synthetic textual data, allowing methodology to be tested without using real-world data in test environments.

Most notably, the capability of LLMs to quickly process a vast amount of textual information and interact with humans in natural languages has a potential to greatly enhance the user experience on statistical dissemination platforms. Currently, the dissemination platform of most statistical organizations is structured by domains and topics. Users need to click through multiple pages, and in a more unfortunate scenario, go through several rounds of back-and-forth, to find the right statistics they are looking for. Also, this structure could be cumbersome for users who seek and integrate data from multiple domains and topics. While statistical organizations have strived to provide products in formats tailored to different audiences (e.g., headline numbers of journalists, raw data for researchers, analysis reports for policymakers),

⁷ Generic Statistical Business Process Model (<https://statswiki.unece.org/display/GSBPM/>).

users who are not familiar with the ways how these can be accessed on the website might encounter difficulties. LLMs can help mitigate these challenges and help improve the quality of data provision to users – the ultimate goal of official statistics producers – through, for example,

- Interactive queries: Enabling LLMs to engage in a dialogue with users to clarify their information needs and refine queries can result in more accurate and relevant responses (see use case in Section 3.3. StatGPT (International Monetary Fund)).
- Customized information delivery: Statistical organizations can allow users to tailor how they receive statistical information from LLMs. Some users may prefer summarized reports, while others may seek in-depth analyses or raw data.
- Data interpretation assistance: LLMs can help users interpret complex statistical data by providing explanations, visualizations, and context. This aids users in understanding the significance and implications of the statistics they are querying.

2.4. Changes the way people find information and knowledge

Statistics organizations have adapted to the changing landscape of information dissemination by diversifying their channels to reach data users and audience as much as possible. Over the past decade, the way people find information has evolved significantly. They rarely directly visit the websites of statistics organizations for official statistics, people often begin their search on platforms such as Google.

These search engines and digital platforms employ algorithms (e.g., Google’s search index), to sift through the vast expanse of information on the web and present users with relevant information. For example, when searching for the “inflation rate of country X in year Y”, these platforms may display the official statistics from the relevant national statistical organization but can also include data from other sources. While the exact workings of these algorithms remain undisclosed, strategies have emerged to enhance the visibility and exposure of content on these platforms which many statistical organizations have adapted to.

However, with the emergence and growing popularity of user-friendly services built upon LLMs (e.g., ChatGPT), the paradigm of information retrieval once again starts shifting. It is already possible for LLMs to retrieve historic statistics from their training data via user prompts without the aid of official statistical organizations. However, there will be data timeliness and quality issues in the output produced, based on the age and source of the training sets used by the LLMs. Timeliness and accuracy issues may not always be obvious to the average user of LLMs, nor may it be obvious that LLMs cannot currently produce up to date meaningful statistics.

While acknowledging the risks of LLM usage, official statistics organizations should understand the capabilities that LLMs offer and potential impacts on the provision of official statistics and traditional statistical use cases.

In order for official statistics to stay relevant in the age of LLMs, statistical organizations should provide services that LLMs cannot do by themselves alone, providing high quality, accurate and timely statistical “source of choice” options for official statistics users.

Official statistics organizations can choose to do this within their own country or organization, or can work jointly together, and with LLM providers, to provide combined statistical products not available today using the power of LLMs. LLMs should be seen as a key enabler for more timely and efficient future provision of statistics, both nationally and internationally.

5. Considerations as statistical organizations move forward with large language models

LLM offers many opportunities for statistical organizations, but it is crucial to proceed with caution while taking various factors into account when integrating LLMs within the

organizations. In this section, we review the main considerations involved in exploring LLMs such as governance, engagement with technology companies, open access models, and public relations. Although the topic is evolving fast, we aim to provide brief practical suggestions at the end of this section.

5.1. Governance

To gain the benefits promised by LLMs/GPTs as outlined in Sections 2 and 3, agencies must put in place new governance measures or integrate their own internal governance framework to limit the risks outlined in Section 4. The risky areas discussed therein include ethics and bias, accuracy, privacy and security, copyright litigation and legal issues, and potential misuse due to lack of literacy and understanding. Potential mitigation strategies were outlined there.

In this section, we consider how we can govern LLMs through implementing these mitigation strategies, in the context of modern statistical agencies operating in an environment already determined by national laws, international frameworks and agreements, existing and changing technical landscapes with dominant players, and existing agency culture.

Governing Large Language Models

Where governance will apply to an implementation or use of a LLM, project stakeholders should establish reasonable and appropriate objectives for the project, aligned with core values of the agency and principles of official statistics and within the national context. We note that governance will always be limited by the fact that the most powerful LLM/GPTs are ultimately owned and controlled by third parties, and due to their size, most often must run on third-party cloud platforms that are also externally controlled.

Therefore, the recommendation is not to implement Responsible AI full track but rather insist on the challenge and conflict generative AI (LLM services in particular) raises with respect to Responsible AI. When adopting LLM/GPTs in organizational workflows (whether as part of third-party Off-the-Shelf (OTS) products, via an API call, or through fine-tuning a foundation model and embedding it in an internally developed and deployed product), we must consider the challenges and conflicts in use of the LLM/GPT in the intended workflow/application and identify appropriate mitigation actions.

Governing large language models / generative pre-trained transformer in current technical landscape

LLMs and GPTs are rarely trained entirely on local or otherwise publicly available datasets. They are often trained, hosted and run on a third-party platform, such as those provided by Amazon (AWS), Google (GCP) or Microsoft (Azure). Agencies will set up agreements with technology vendors to ensure key national interests are protected and relevant laws are adhered to (e.g., keep data hosted on local servers). However, it will remain the case that some parts of any LLM/GPT pipelines and products used by agencies will not be in our control, and further, may not even be entirely visible to agency staff.

Therefore, the nature and level of governance of LLM/GPTs within statistical organizations will depend on how the LLM/GPT is entering into the organization's sphere. Governance of a project where an LLM/GPT is being developed (e.g., fine-tuned) will be different to governance surrounding use of a third party closed-source application. For each case, governance will require outlining the risks and specifying appropriate mitigations. Further details of classes of risk and potential mitigations are articulated in Section 4.

Some examples of governance of LLM/GPTs are given below.

Example A: A Licence Agreement to Install LLM/GPT-based Third Party Application

Microsoft will embed its CoPilot AI tool in its Office365 suite, which it claims is expected to improve workplace productivity. Some level of governance will occur at the legal level – e.g., the requirement that data be hosted onshore. However, some governance will need to be addressed through softer measures once CoPilot is installed and in use. For example,

statistical organization staff who query the AI-assisted tool for information, may be overly confident in the accuracy of the output, and publish/communicate potential misinformation, or make decisions based on incorrect or incomplete information. For further discussion, see the example in Section 3.1 regarding errors discerning human vs LLM-generated occupation task list generation, and the general discussion around misuse in Section 4.5. Statistical organizations cannot eliminate risk of misuse but can put in place mitigations outlined in Section 4.5 around improving data and AI literacy, and establishing clear protocols for use, and insertion of technical guardrails preventing misuse.

Example B: An internally developed pipeline or product which makes use of a pre-trained LLM/GPT

Increasingly, the trend is for internal staff developers who are familiar with agency goals, datasets, and use cases, such as data scientists or machine learning engineers, to use pre-trained models (also called foundation models). The agency will be limited in its ability to fully govern the product or pipeline which makes use of the foundation model.

For example, it will be hard for statistical organizations to ensure the product or pipeline does not use components (datasets or code) labelled or developed in environments practising poor human labour standards. It will be hard to prove data accuracy is acceptable and the model is unbiased as outlined in Section 4.2, or that data poisoning has not occurred as outlined in Section 4.3.

Even when the third-party makers of those foundation models release training code or training data through a public repository and/or offer users a less restrictive open-access or open-source licence, there is still a lack of transparency. Indeed, the Foundation Model Transparency Index released by the Stanford-based Center for Foundation Models scored many prominent Foundation Models out of 100, awarding a point for each criteria where the company provided sufficient information to each question. Meta's Llama 2 model received the top score of 54/100. That means in 46 criteria, Meta did not provide sufficient information for the researchers to consider that transparency criterion to be satisfactory.⁸

Given these conflicts and tensions, we do not recommend banning LLM/GPTs as this creates a risk of shadow AI with statistical organizations. Rather, we recommend assessing each project or application for risks, and putting in place appropriate mitigations.

Governance in effect – evaluation and monitoring

Evaluation metrics: Where an LLM is used to provide answers to queries or recommendation, the LLM performance should be evaluated for criteria such as faithfulness (e.g., is the generated text faithful to the source document?), reproducibility (e.g., does it return the same or similar outputs for the same or similar query) and relevance (does the response answer the query?). Evaluation also covers how the outputs reflect organizational values (e.g., might the returns lead to reputational damage?). Developers might also need to consider adjusting parameters so that the tone of outputs are unbiased, politically neutral and factual, and that outputs are aimed at the appropriate audience (whatever that audience might be). Text generation outputs should be checked to ensure these are not unintentionally plagiarizing existing publications – while computer-generated text are still to be finalized, the negative publicity and possible impact on public trust is not worth risking.

Monitoring: The degree to which an AI/LLM pipeline or product meets the objectives or raises risk should be measured, monitored and reported correctly during the lifetime of the pipeline or product. A monitoring step could be for stakeholders to perform threshold or impact assessments, where project development and product use is scored against relevant risk categories. In order to ensure AI systems remain responsible over time, development should include a maintenance plan, including how often training data will be refreshed, and methodological and code reviews to ensure the AI model is up-to-date. The points above should be integrated into maintenance to account for changes to each of these over time.

⁸ <https://crfm.stanford.edu/fmti/>.

5.2. Engagement with tech companies who provide large language model services

The LLM ecosystem is a complex and rapidly evolving field. Central to this ecosystem are major entities like Google, OpenAI, Microsoft and Meta AI, which play a pivotal role in defining and advancing LLM technologies. Within this context, it is vital for statistical organizations to also explore and emphasize the use of open-source models and platforms. Companies such as Hugging Face and EleutherAI, which are built on open-source ideologies, contribute to creating a more diverse and accessible environment. Engaging with these entities requires balancing proprietary and open-source technologies to drive innovation and maintain ethical standards.

Understanding the diverse roles of technology companies in the LLM ecosystem is essential. By considering factors such as primary offerings, roles within the ecosystem and the range of services provided, statistical organizations can effectively navigate this space.

Role of cloud providers

Cloud service providers are integral to the operation and advancement of LLMs. When engaging with these providers, statistical organizations must consider several key factors. Data privacy and security are paramount, as is the scalability and performance of the services. Cost management is another critical area, requiring a clear understanding of pricing models and potential hidden fees. Ensuring legal compliance, such as service availability in specific regions (e.g., Europe or Western Europe) and technical compatibility with existing systems of a statistical organization are also crucial considerations.

The global cloud market is primarily dominated by the ‘big three’: Azure, AWS and Google Cloud. However, alternative providers often specialize in niche services that offer specific integrations, potentially more suitable for certain statistical organizations. Selecting a cloud provider for AI infrastructure or platforms necessitates aligning with specific needs and considering longer-term development paths. Being mindful of the risks of dependency on major key players in the LLM ecosystem is also important.

Large language model ecosystem

In the LLM ecosystem, the services offered by tech companies often span multiple categories, highlighting the interconnected nature of this field. For instance, Azure Machine Learning by Microsoft allows users to access models developed by OpenAI and Meta AI and some of the models at Hugging Face. Similarly, Hugging Face distinguishes itself by offering a wide array of services across nearly all categories in the LLM ecosystem.

For statistical organizations, recognizing and understanding these multifaceted roles is crucial. By identifying the specific category or categories a company operates in, statistical organizations can more effectively strategize their engagement with tech companies. This knowledge allows them to pinpoint which companies offer the most relevant and beneficial services for their particular needs, whether it is for leveraging advanced AI models, accessing diverse datasets or utilizing efficient training platforms. Furthermore, understanding these categories helps statistical organizations anticipate and navigate potential overlaps in services and collaborations, ensuring a more streamlined and efficient approach to integrating LLM technologies into their operations.

LLM developers and providers category includes companies specializing in the research, development, and deployment of LLMs. Notable examples include OpenAI, Meta AI, Google DeepMind, as well as open-source players such as EleutherAI and the Technology Innovation Institute (TII). These organizations are at the forefront of advancing LLM technologies. From the technical point of view, ensuring that the LLMs from these providers can be seamlessly integrated into the systems of statistical organizations is crucial. This involves compatibility with existing infrastructure and the ability to adapt to specific technical requirements. Alignment with the ethical standards of statistical organizations is paramount. It is essential that the LLMs adhere to principles of responsible AI, including transparency, fairness, privacy, and accountability. Ensuring that these models are developed

and deployed in an ethical manner aligns with broader societal values and regulatory frameworks.

AI infrastructure and platform providers is the second category in the LLM ecosystem, and it includes companies that provide the necessary hardware and software infrastructure to train, deploy and run LLMs, such as Microsoft Azure ML, Google Cloud AI platform, AWS SageMaker and more. For statistical organizations, engagement with these providers necessitates a focus on scalability, performance, technical compatibility, and a thorough understanding of the cost structures, including any potential hidden expenses.

LLM application developers are tech companies who are instrumental in developing applications or services that utilize LLMs for specific functionalities like chatbot development. The innovation in application development, user-centric design and adherence to data privacy standards are vital aspects of their contribution.

AI customization and fine-tuning services is a crucial segment of companies that includes AI startups and specialized technology firms that tailor existing LLMs to meet specific customer needs. Their adaptability and ability to integrate customized solutions into existing systems are key considerations for statistical organizations.

Equally important are the **LLM research and innovation labs**, which include academic research labs and R&D departments. These entities push the boundaries of what LLMs can achieve, focusing on cutting-edge research and ethical AI practices. Their work significantly contributes to the broader AI and LLM knowledge base. Engagement with these labs can provide statistical organizations access to the latest research and ethical AI practices.

In the ecosystem the **LLM community and open-source initiatives** play a pivotal role. Platforms like Hugging Face and various GitHub repositories dedicated to LLM research foster community engagement, promoting an open-source culture in LLM development. These initiatives drive innovation and ensure the accessibility of tools and resources, crucial for a collaborative and inclusive LLM ecosystem. Statistical organizations ought to collaborate with these initiatives to gain access to a rich array of open-source tools and resources.

In the very last category, **data and training services for LLMs**, there are companies that are essential in supplying the vast and varied datasets required for training LLMs. These entities not only provide data, but they can also offer crucial services that facilitate the training process of LLMs. Companies like EleutherAI and Hugging Face stand out in this domain, offering a range of datasets and tools that are vital for the development of robust and effective LLMs. Their contribution is crucial in ensuring that LLMs are trained on diverse, extensive, and high-quality datasets, which is fundamental for the accuracy and reliability of these models. Additionally, these services often include tools and platforms that assist in the efficient and effective training of LLMs, making them an indispensable part of the LLM ecosystem. Statistical organizations should engage with these entities for high-quality, diverse data sources and efficient training platforms.

Each category within the LLM ecosystem offers unique opportunities for engagement, contributing to the overall growth and ethical use of LLM technologies.

5.3. Considerations around open access

There are several aspects about LLMs and providers to be used by a statistical organization that need careful consideration. The main dimensions to be considered encompass the accessibility of the model's underlying structure and training data, the licensing terms governing the model's utilization, and the access to inputs and outputs when utilizing the LLM. The evaluation typically involves a trade-off analysis, balancing the benefits of convenience against the need for control. Cost and access to skills are also relevant points of consideration, as LLMs require significant IT infrastructure and expertise for their operations at scale.

The accessibility of LLMs spans a wide spectrum. Some models are openly accessible, allowing inspection and modification of their architecture and weights through fine-tuning.

Conversely, others are maintained as proprietary assets, with access only granted via APIs or other interfaces. In certain cases, providers of closed-access models may still offer options for fine-tuning, permitting users to adapt the model's weights to their specific data and use cases. While direct access to model weights may appear inconsequential, the capacity to customize an LLM to particular data and use cases can prove highly relevant for a statistical organization.

Regarding openly accessible models, diligent examination of the licensing terms is imperative. LLM creators may impose specific conditions governing the utilization of the model, which may be inadvertently breached by uninformed users.

Transparency and accessibility to the model's training data are paramount for assessing the potential presence of biased, harmful, or copyrighted material which may influence the output generated by the model (as discussed in Section 4). In such cases, complete transparency and access are indispensable to mitigate any reputational risks, given that the training dataset significantly shapes the model's output. Another aspect of consideration regards the accessibility of LLMs to contemporary information. The knowledge corpus employed in model training is delimited by a cut-off date predating the start of the model training process. Addressing this limitation entails the integration of updated content, a practice often referred to as Retrieval Augmented Generation (RAG). Furthermore, certain frameworks are equipped with mechanisms enabling LLMs to access real-time data from the internet.

Finally, the issue of confidentiality concerning input and output data merits careful consideration. In many cloud-based services, both the input and output data may be retained by the service provider to facilitate future LLM iterations through training and fine-tuning. Consequently, users may encounter limitations regarding the use of confidential information. However, it is worth noting that some vendors are beginning to provide access to closed models in a sandboxed environment, offering users the ability to maintain full control and privacy over inputs and outputs.

In summary, statistical organizations should assess the potential benefits and disbenefits of open access models when evaluating LLMs, in particular the level of transparency and ability to collaborate with other statistical organizations given the open nature of these models.

5.4. Communication with public

The field of LLMs is fast-evolving. While these models offer astonishing capabilities, their rapid development places this AI technology in a rather grey zone where public opinion and sentiment can be uncertain and prone to shifts.

As a government agency whose products significantly influence policy and decision-making with national impact, statistical organizations bear a great responsibility to use the LLMs in a responsible way as well as communicate it in a transparent way to the public. The very fact that statistical organizations' core business (i.e., production of official statistics and data services) heavily relies on the public trust requires that statistical organizations should pay even more attention and invest in communication to society, in particular, for data providers who could raise concerns that their data may be misused while statistical organizations are interacting with LLMs. After all, the public has a fundamental right to understand how their data might be utilized and to be assured that measures are in place to protect their data.

In communicating the utilization of LLMs, it would be important to convey that statistical organizations are:

- using LLMs purposefully where there are clear benefits: it is essential to clearly communicate why LLMs are used in statistical organizations and highlight the tangible benefits of this technology (e.g., increased efficiency, cost savings, improved services), with concrete examples where the use of LLMs has yielded success. For example, LLM-based chatbots that help the public better understand and access statistical data is one of the roles that generative AI/LLMs can play quite autonomously.

- aware of limitations and risks: it is important to demonstrate that statistical organizations are not using LLMs blindly and aware of the potential limitations and risks associated with LLMs. Areas that LLMs are and will not be used (e.g., for making individual predictions that could adversely affect people) could be mentioned.
- taking necessary mitigation measures: it is vital to explain the steps taken to mitigate the limitations and risks (e.g., measures taken to maintain data confidentiality and security) while emphasizing that human intervention is in place to oversee and guide the use of LLMs.

In terms of internal communication, it would be important to consider what a particular use case might unintentionally say about organizational priorities to its employees. For example, LLMs might provide an efficient way of producing non-technical summaries, but this could also be seen as an organization outsourcing this task to a model rather than fostering skills internally in non-technical writing to communicate with interested members of the public.

5.5. Practical suggestions and concluding remarks

The use of LLMs by statistical organizations is still in its infancy, and the landscape is evolving fast. Best practices are being developed over time and will require constant effort from statistical organizations to keep up to date. There are a few practical suggestions that we feel are relevant in the short term and may stand the test of time.

The first one is to provide training on LLMs at all levels in the organization – technical, operational, and managerial – to raise awareness and better understand LLMs capabilities and limits.

Secondly, we would suggest approaching LLMs with the execution of small pilot projects to gain familiarity with the technology and understand the potential value that could be generated. Those small-scale projects may be able to ramp up the capabilities of statistical organizations on the subject, deliver results that could justify and guide further investments, and ultimately mitigate the risks of exploring the use of LLMs.

Thirdly, statistical organizations should develop an overall LLM strategy once awareness and familiarity are at sufficient level, having completed some small-scale projects as discussed above.

Finally, statistical organizations should devote continuous effort to keep up to date with the continuously changing landscape of LLMs, both from a technological and strategic point of view.

Recognizing the swift advancements in LLMs, we understand that the pace of progress is beyond our complete understanding. This white paper aims to collect existing use cases up to the present day and deeply explore the topic from different angles relevant to statistical organizations. Due to the dynamic nature of this field, working together is crucial. Therefore, we invite experts to collaborate, share insights, and collectively navigate this ever-changing landscape. Our dedication to explore this topic continues, and we welcome ongoing participation in this exploration.
