Large Language Models for Official Statistics

HLG-MOS White Paper

December 2023

Foreword

Late 2022, ChatGPT 3.5 took the world by storm. The news of its ability to create good-looking text, based on a natural language request from a user, quickly spread. All over the world people from quite different backgrounds started experimenting and shared their experiences with friends and colleagues. When the first wave of euphoria had calmed down a bit, people realised some aspects of ChatGPT might require critical examination. Moreover, other Large Language Models (LLMs) had appeared that required attention. All this led to more systematic and organised activities in specific communities.

The worldwide official statistics community was no exception. From early 2023, in many of the meetings of the community the topic of ChatGPT was raised and participants wanted to share views and experiences. Two groups under the umbrella of the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) then joined forces, the Blue Skies Thinking Network and the Modernisation Group on Advancing Data Science and Modern Methods. They organised a series of (mostly online) meetings to discuss aspects of ChatGPT and other LLM. Mid 2023 it was decided to bring a group together to draft a white paper on this topic. A draft of the white paper was presented at the annual HLG-MOS workshop in Geneva (November 2023). Moreover, at this workshop it was decided to launch a formal HLG-MOS project on Generative AI including LLMs in 2024, which is open to active participants from the official statistics community, and supporters from the research world and other domains, who have an interest in and enthusiasm for this topic.

The white paper is formally released in December 2023. It provides a broad overview of LLM aspects and ideas relevant from an “official statistics” perspective - although much of the content may be relevant to a wider audience. It has been produced by a large and diverse team of writers, who have done a great job in a short time. Of course, the white paper reflects the current state of the art and in the dynamic LLM world, its content may quickly become outdated. Therefore, the 2024 HLG-MOS project on Generative AI will be tasked to review and build upon the white paper to help the collective knowledge of the community remain up-to-date. Hopefully the white paper finds its way to many readers!

On behalf of the HLG-MOS community,

Barteld Braaksma (Statistics Netherlands), Chair of the Blues Skies Thinking Network

Gary Dunnet (Statistics New Zealand), Chair of the Modernisation Group on Advancing Data Science and Modern Methods

Acknowledgement

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.

However, as it often happens with new technology in its early stage, each organisation 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 organisations invaluable to facilitate the adoption of the new technology.

To respond this challenge, the UNECE High Level Group on Modernisation of Official Statistics (HLG-MOS) Modernisation Groups - Blue-Skies Thinking Network (BSTN) and Applying Data Science and Modern Methods (ADSaMM) Group initiated to draft a white paper on LLMs in the context of official statistics.

This publication is a result of a 4-month journey of a team of experts from various national and international statistical organisations around the world to provide a common understanding of implication of LLMs for official statistics.

The following team members kindly dedicated their time, and contributed their knowledge, experience, and expertise[*]:

- Cathal Curtin (Lead Editor) and Pubudu Senanayake – New Zealand (Statistics New Zealand)
- Claire Clarke, Ilana Lichtenstein and Andrew Jamieson – Australia (Australia Bureau of Statistics)
- Shirin Roshanafshar and Wesley Yung – Canada (Statistics Canada)
- Riitta Piela - Finland (Statistics Finland)
- Vytas Vaiciulis– Ireland (Central Statistics Office)
- Andrea del Monaco and Luigi Palumbo – Italy (Bank of Italy)
- Vera Toepoel – The Kingdom of Netherlands (Statistics Netherlands)
- Karen Tingay – The United Kingdom of Great Britain and Northern Ireland (Statistics Authority)
- Andrew Banks– The United Kingdom of Great Britain and Northern Ireland (Office of National Statistics)
- Bilyana Bogdanova, Olivier Sirello and Krzysztof Zdanowicz – BIS
- Jean-Marc Museux and Cristiano Tessitore - Eurostat
- Jeff Danforth and Jim Tebrake - IMF
- InKyung Choi and Amilina Kipkeeva - UNECE

[*] The views and the ideas expressed are those of the authors and do not necessarily reflect those of their respective institutions.

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 organisations' operations into the future. Like any offices in many sectors and domains, statistical organisations 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 works, 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 organisations 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 organisations 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 organisations 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 organisations, communicating the responsible use of LLMs – that statistical organisations are using them purposedly where there are clear benefits with awareness of risks and necessary mitigation measures – is vital. The use of LLMs by statistical organisations is still in its infancy, but there are a few practical suggestions:

- provide training on LLMs at all levels in the organisation (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 organisations will continue to be crucial to collectively explore different applications and share insights and experiences along their journey.

Table of Contents

Acronyms

ABS	Australian Bureau of Statistics
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BIS	Bank for International Settlements
CoP	Community of Practice
CPI	Consumer Price Index
CSO	Central Statistics Office
EU	European Union
GANs	Generative Adversarial Networks
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Units
GSBPM	Generic Statistical Business Process Model
HLG-MOS	High-Level Group for the Modernisation of Official Statistics
HR	Human Resource
IMF	International Monetary Fund
IP	Intellectual Property
IT	Information Technology
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Machine Learning
MLSecOps	Machine Learning Security Operations
NLP	Natural Language Processing
NLU	Natural Language Understanding
NSO	National Statistical Organisation
OTS	Off-the-Shelf
RAG	Retrieval Augmented Generation
SDMX	Statistical Data and Metadata eXchange
SMA	Subject Matter Areas
VAEs	Variational Autoencoders

INTRODUCTION

01

1. Introduction on LLM

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 summarisation.

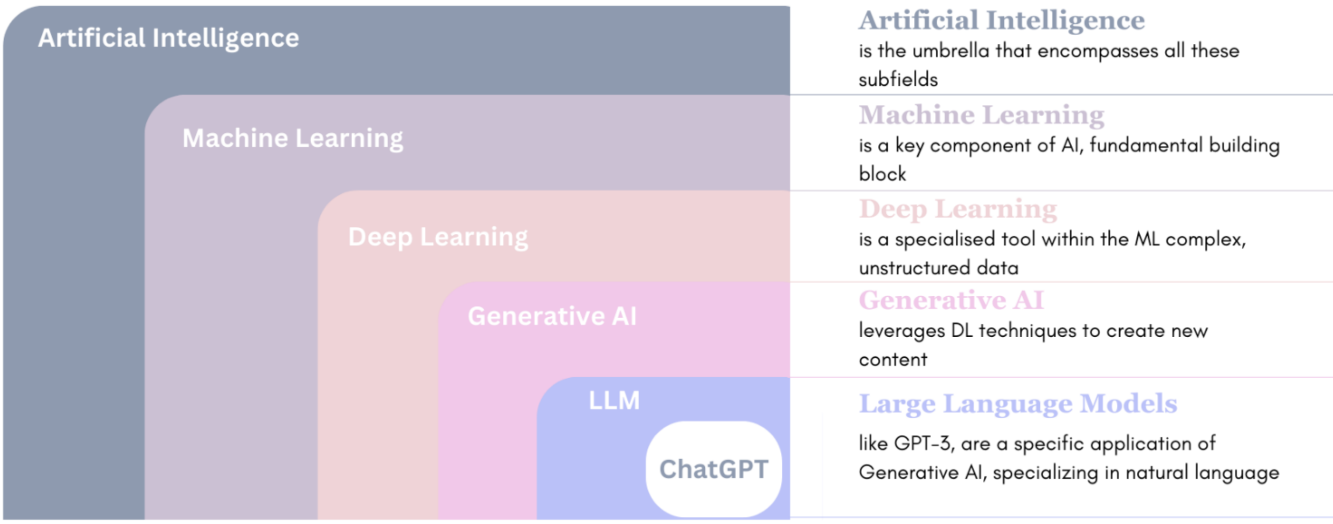
In services developed based on LLM (e.g., ChatGPT), users can interact with LLMs through natural languages, called "prompts" (instruction that generates responses from LLMs), for example, as below:

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 AI, ML and Generative AI

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[2].

- **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 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 specialised 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).



[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8259629

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 models[3] 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 optimisations 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 minimise 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 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

[3] 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 sets them apart from traditional machine learning models. They can be used as a basis for the development of specialised subsequent applications.

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 generalise 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 customise 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 capitalises 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 specialised 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 specialised 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 organised 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[4]. “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 customise 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 check carefully 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.

[4] https://opensource.com/resources/what-open-source

IMPLICATION AND OPPORTUNITIES FOR OFFICIAL STATISTICS

02

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 organisations, from communication to project management, highlighting their role in optimising 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 Organisations Can Do and Cannot Do with LLMs

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 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 organisations. They are very good at understanding textual information, summarising large amounts of information, and generating human-like responses which could be useful in automating many tasks within a statistical organisation. This section will present some ideas where LLMs could be used, including tasks that are needed in any organisation such as drafting emails and preliminary reports, summarising information for brainstorming sessions, project management and translation to multiple languages. Also tasks that are particularly relevant for statistical organisations such as text classification, data visualisation 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 organisations 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 organisation, statistical organisations 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 organisation to function effectively, they require a significant amount of time and effort from dedicated staff.

LLMs/ChatGPT can help streamline operations within statistical organisations 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 organisation 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[5]. This saves time and increases the quality of written materials. For reports, LLMs summarise lengthy documents, provide options for data visualisations, identify errors, and offer recommendations. They can also generate customised report sections and automatically create tables, charts, and lists. However, it's essential to avoid using sensitive data, as outlined in risks related to LLM use discussed in Section 4.
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, optimising 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 summarised. Additionally, LLMs facilitate meeting management by automating the creation of meeting agendas and suggesting discussion topics in line with predefined goals or recent updates. In Q&A session preparation it can be useful for compiling relevant questions aligned with the meeting's agenda and topics.
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. The choice between LLMs and automatic systems depends on specific translation needs and priorities.
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 customisation of style, structure, and slide quantity, but also for developing talking points for a more human-friendly presentation tone. Another useful application of the LLMs for presentations is a customisation for different audiences, simplifying complex concepts and avoiding technical terms when needed.
6. **Educational purposes** - LLMs can be employed for educational and training purposes within the organisation. They can provide explanations, create quizzes, and assist in designing e-learning materials to enhance the skills and knowledge of the workforce. It can also create tailor-made timeframes, set deadlines, and act as a "personal tutor".

[5] Note that LLM generated emails may be easy to recognise; 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

7. **Image Generation** - Stock imagery is often used with reports and productions of statistical organisations. Rather than purchase stock imagery, statistics organisations 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.

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 organisations, for example,

- Design collection (GSBPM[6] 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 organisations can use them for organising 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 preprocessing 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 organisations:

- Assist coding and translating between programming languages: LLMs can deal with not only natural languages but also programming languages which statistical organisations 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 optimising code development, providing code snippets and translating between different

[6] Generic Statistical Business Process Model (https://statswiki.unece.org/display/GSBPM/)

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 organisations 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 organisations have strived to provide products in formats tailored to different audiences (e.g., headline numbers of journalists, raw data for researchers, analysis reports for policy makers), 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))
- Customised information delivery: Statistical organisations can allow users to tailor how they receive statistical information from LLMs. Some users may prefer summarised 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, visualisations, 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 organisations 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 visit directly the websites of statistics organisations 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 organisation 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 organisations 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 organisations. 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.

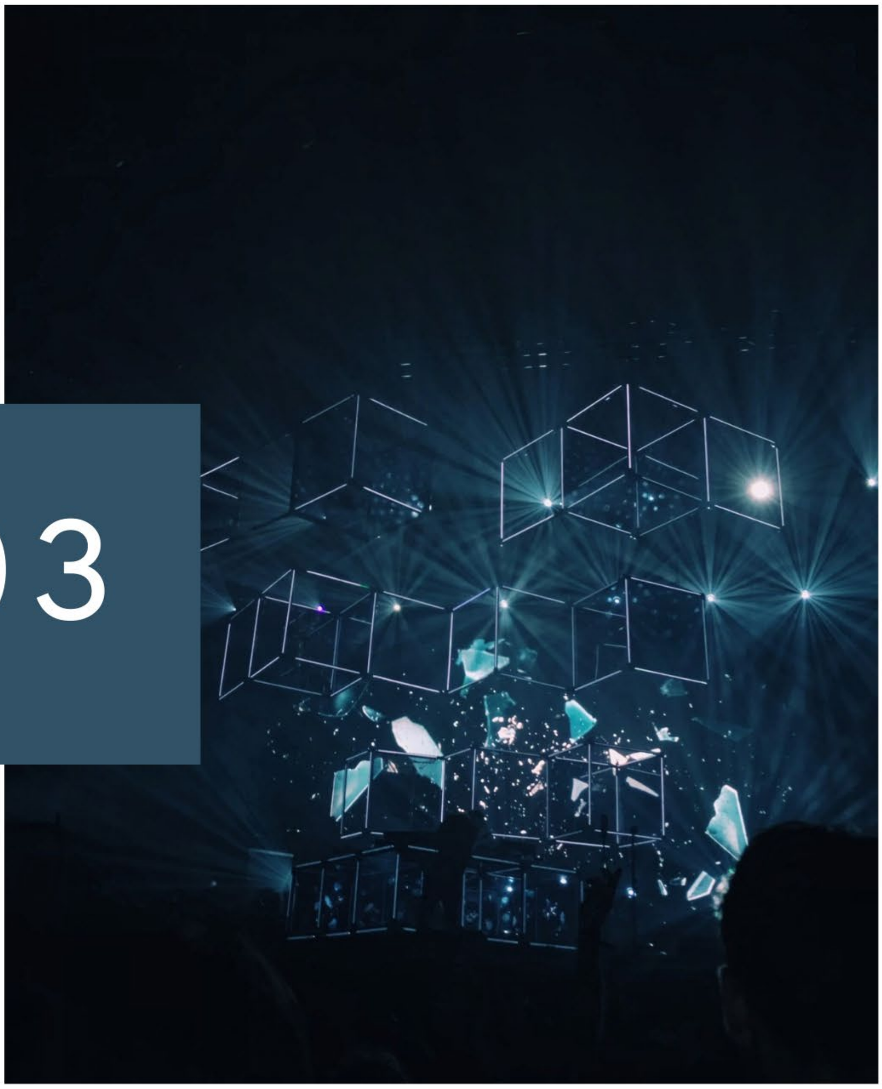
While acknowledging the risks of LLM usage, official statistics organisations should understand the capabilities that LLMs offer and potential impacts on the provision of official statistics and trial statistical use cases.

In order for official statistics to stay relevant in the age of LLMs, statistical organisations 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 organisations can choose to do this within their own country or organisation, 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.

USE CASES FOR STATISTICAL ORGANISATIONS

03



3. Use Cases for Statistical Organisations

Despite the rapid development of LLM, some statistical organisations have already tested it in their works and even launched several product updates. In this section we present implementation examples on different use cases, each of which highlights the opportunities, challenges and risks associated with LLM implementation. Each use case is presented in a structured format detailing the business problem solved, the value added by the LLM solution, a description of the solution, preliminary results, IT and human resource requirements, validation processes, stakeholder involved, organisational barriers, risk mitigation strategies, collaboration within and outside the organisation, the current state of implementation, possible extensions of the use cases and suggested next steps.

3.1. Updating Statistical Classification Definitions (ABS)

Business Problem and Value Addition

The Australian and New Zealand Standard Classification of Occupations (ANZSCO) [7] provides the basis for the standardised collection, analysis and dissemination of occupation data for Australia and New Zealand. The Australian Bureau of Statistics (ABS) is undertaking a comprehensive review of the classification to reflect the contemporary labour market and better meet stakeholders' needs. As part of the comprehensive review, all occupations within ANZSCO are being updated to include a list of tasks performed by that occupation.

Creating task lists for occupations is a time consuming process, requiring substantial research and review of stakeholder submissions. This process can take between one hour and multiple days depending on the complexity of the occupation.

The ANZSCO Review team sought to provide an efficiency boost to this process by using LLMs to generate an initial set of tasks for each occupation under review. These task lists provide analysts with a starting point, which they can review and edit as appropriate.

Preliminary Results

The ANZSCO review is being conducted in multiple tranches covering differing numbers of occupations. Tranche 1, which covered approximately 160 of ANZSCO's 1,076 occupations, was conducted without the use of LLMs. This provided a set of analyst generated occupation task lists for these occupations that could be used as a test dataset to assess the quality and usefulness of LLM generated task lists according to the methodology set out in Quality Metrics.

The project board made a go/no go decision to continue the project and provide analysts with LLM generated task lists for occupations covered under tranche 2, assessing that the LLM was producing task lists that were of suitable quality to be useful to analysts and that they would afford time savings across tranche 2. A set of task lists was then produced covering over 400 occupations.

[7] Introduction | Australian Bureau of Statistics (abs.gov.au)

IT and HR Requirements

HR Requirements

Implementation of this project required contributions from the ANZSCO Review team; and ABS legal, methodology, policy and technical services; for a combined estimated workload of approximately seven Full Time Equivalent (FTE) work weeks.

IT Implementation

Delivery of the occupation task lists involved producing a relatively simple python script run on a desktop computer to query the LLM API. The prompt used to query the LLM was revised and optimised as per the quality metrics discussed in Quality Metrics.

Quality Metrics

Definition of Quality Metrics Used

A critical issue for the approval of the project was measurement of the quality of the LLM generated task lists.

As discussed in Preliminary Results, the provision of analyst generated task lists from tranche 1 of ANZSCO Review allowed for the quality of LLM generated task lists to be calculated by comparing them to the analyst generated task lists.

Precision and recall were calculated based on a modified BERTScore[8] methodology. Reasons for altering the BERTScore calculations include:

- BERTScore is designed to calculate the precision and recall of words in pairs of sentences using word level embeddings. This use case requires calculating the precision and recall of the semantic content of sentences within batches of sentences.
- High quality sentence level embeddings that enabled efficient calculation of semantic similarity of entire sentences were not available at the time BERTScore was proposed.

The calculations of precision and recall were:

$$P(c,a) = \frac{1}{|c|} \sum_{c_i \in c} (c_i, a_j)$$

$$R(c,a) = \frac{1}{|a|} \sum_{a_j \in a} (c_i, a_j)$$

where **c** represents the set of vector embeddings of LLM generated tasks and **a** represents the set of vector embeddings of analyst generated tasks.

[8] 1904.09675.pdf (arxiv.org)

An additional metric denoted "batch similarity" was calculated as:

$$Batch\ Similarity(a_b, c_b) = cosine\_similarity(a_b, c_b)$$

where c_b represents the vector embedding of the entire batch of LLM generated tasks for an occupation and a_b represents the vector embedding of the entire batch of analyst generated tasks for the same occupation.

For all calculations above, vectorisation was performed using the mpnet_base_v2 SBERT model.

Quality Metric Results

Table 3.1 shows the highest average quality metrics achieved after testing multiple LLM models with multiple prompts.

Table 3.1: Quality Metrics for ANZSCO Review LLM Use Case

Average Precision	Average Recall	Average Batch Similarity
0.69	0.69	0.85

This aligned to team expectations that a quality value of approximately 70% is considered reasonable in applications of machine learning.

Stakeholder Response

At the time of writing, no official data had been collected on the efficiency improvements afforded by this project. In lieu of official metrics, multiple ANZSCO Review team experts were asked to comment on the quality of the LLM generated task lists and on the estimated time savings afforded by being provided them.

The feedback on the quality of the LLM generated task lists was that they were "good and relevant to the occupations albeit slightly generic in some cases". Estimates on time savings varied but averaged to approximately 2 hours per occupation, which, if accurate, will result in 1600+ work hours saved over the life of the ANZSCO Review project.

An additional informal yet amusing and insightful test was conducted in which analysts and senior executives were shown an LLM generated task list and an analyst generated task list for the same occupation and asked to determine which was which. Approximately two-thirds of participants were incorrect, again indicating the LLM generated tasks were high quality.

As the review progresses, analysts are continuing to request additional task lists for new and revised occupations within ANZSCO. This indicates analysts have a high level of confidence in and acceptance of LLM as a technology to assist their work:

- Measuring the quality of LLM outputs prior to analysts accessing the results grants confidence at all levels of management that using LLMs is worthwhile; and
- Strong project governance maintaining multiple layers of human centred review emphasises the ongoing importance of analysts and therefore limits anxiety of jobs being replaced by AI.

Barriers

Approval to conduct this project was subject to the following requirements within ABS:

- Development of a quality metric and target approved by the methodology experts;
- Legal advice on the use of LLMs;
- An assessment of the project against appropriate Australian ethics frameworks;
- Establishment of strong project governance; and
- Provisioning of a technical capability of accessing an LLM API.

Risks and Mitigation

Risks identified in the application of LLMs for this project are shown in Table 3.2.

Table 3.2: Risks Identified in ANZSCO Review LLM Use Case

Risk Category	Description	Mitigations
Reputation	If the public believes that a National Statistical Organisation (NSO) is using an AI for decision making, there is a risk to the organisation's reputation and social licence to operate.	ABS committed to: • Ensuring the use of LLMs adhered to multiple suitable ethical frameworks;[9][10] • Ensuring the review process contained multiple layers of human centred review of LLM generated content; • Making a public disclosure about the use of LLMs in ANZSCO Review and the nature of the human centred review process.
Quality	The LLM generated output may not be suitable quality for release.	• Evaluate quality of LLM generated outputs prior to providing to analysts; • Ensure multiple layers of human centred review.
Security	Use of LLMs may present multiple cyber security risks including (but not limited to): • Facilitating malware to produce code that bypasses standard firewalls and	Appropriate protections for cyber security risks raised by LLM technologies are organisation dependent. Appropriate solutions may include: • Firewall restrictions;

[9] Australia's AI Ethics Principles | Australia's Artificial Intelligence Ethics Framework | Department of Industry, Science and Resources
[10] Interim guidance for agencies on government use of generative Artificial Intelligence platforms | aga (digital.gov.au)

	antivirus solutions ● Release of sensitive information within prompts used for the LLM	● Locally hosted infrastructure; ● Running code outside of the corporate environment. This use case cannot inform on the appropriate solution for other organisations. Given ANZSCO Review is concerned with publicly available definitions, it was concluded there was little information security risk in this application.
Legal	The use and publishing of LLM generated material may raise IP concerns.	IP laws surrounding LLMs are largely untested worldwide. The ANZSCO Review team sought legal advice and concluded this was a low risk application. This result cannot be assumed to apply to other projects, other organisations or other countries.

Next Steps and Other Potential Use Cases

ANZSCO Review will continue to apply LLMs to the comprehensive review while also testing newer, higher quality LLMs as they become available. Once the comprehensive review of ANZSCO is complete and ABS moves to an ongoing maintenance phase, there is a risk to ongoing use of LLMs for this purpose. This is presented in Table 3.3.

Table 3.3: Future Risk of Use of LLMs in ANZSCO Review

Risk Category	Description	Mitigations
Quality	Once ANZSCO Task Lists are published in the public domain, LLM providers will use the ANZSCO definitions as part of their training datasets. Continued use of LLMs for this purpose will therefore involve AI trained on AI generated content, which is known to decrease quality.	Ongoing focus on human centred review on AI generated content; Restriction of LLM use cases to generating content that is strictly new.

There is potential for this use case to be extended to reviews of other classifications managed by ABS, including the Australian and New Zealand Standard Industrial Classification (ANZSIC) and the Australian Standard Classification of Religious Groups (ASCRG). The exact applications and desired outputs for these classifications will be developed when their major reviews are conducted.

3.2. Code Translation and Explanation (SAS to R) Using LLMs (CSO Ireland)

Business Problem

The Central Statistics Office (CSO) Ireland has made the strategic choice to switch from SAS to R as its main programming language. This modification demonstrates the company's dedication to utilising R's robust statistical capabilities, open-source ecosystem, and community-driven developments.

The main consequence of this modification is the requirement to translate old SAS code into equivalent R code. This translation is not merely a syntactical exercise; it requires a thorough understanding of the data structures, characteristics, and nuances of both languages. Additionally, because of the sizable library of SAS code that has grown over time, this change necessitates a significant and methodical quality transfer.

Description of Solution and Value Addition

Our data science section at CSO started investigating Large Language Models (LLMS) in the first quarter of 2023. Our aim was to make the conversion of SAS code to R code simpler. We chose OpenAI over ChatGPT because their API offered better control and governance, which are essential for projects requiring precise and specialised solutions.

Our initial platform usage demonstrated the technology's potential for converting SAS code to R. Despite the flaws in certain translations, they provided a strong foundation and drastically reduced the need for manual intervention. There were many new things to learn as we experimented with the various models that were available, each of which had varying degrees of success.

After finding errors in the initial code translations, our team focused its efforts on improving the translation processes. This required making changes to the model's parameters. In order to further improve the calibre of translations, we also started fine-tuning particular models. In this procedure, SAS and R codes were systematically paired. We noticed gradual improvements in the precision and dependability of SAS to R translations as a result of fine-tuning.

As we adjusted the parameters and improved our models, GPT-3.5 and GPT-4 were released. Because of the improved capabilities in these more recent versions, our SAS to R translations are now much more successful. Their quick implementation enabled us to further improve our code translation process.

The GPT-3.5 Turbo, which can be fine-tuned, and the more sophisticated GPT-4, which cannot be fine-tuned at this time, are the two notable models that are now available. The latter excels in contextual comprehension as well as translation jobs, providing useful comments to the translated code.

With GPT-4's introduction, our team swiftly integrated its capabilities into an in-house application (SAS to R Code Assistant). This move allowed us full governance over its deployment and usage, ensuring alignment with our specific needs and objectives.

Key Highlights of the Internal Application (SAS to R Code Assistant):

- **Local Hosting:** By opting for local hosting within our office premises, we've ensured data security and confidentiality. This method reduces the risk of sensitive code or data being exposed externally, ensuring compliance with our stringent data protection standards.
- **Integration with GPT-4:** The application has been seamlessly integrated with the GPT-4 model via the OpenAI API. This allows for real-time interactions, enabling swift translations and potentially other applications like code documentation, analytics insights, and more.
- **Development in R:** In alignment with our strategy to transition towards R, the application has been developed using the R programming language. This not only streamlines the integration of translated code but also makes it easier for our R-centric development team to maintain and upgrade the application.
- **Shiny Application:** Hosting it as a Shiny application brings forth an interactive web interface that is user-friendly. With Shiny, we're able to provide our developers and analysts with a dynamic platform where they can input SAS code and instantaneously receive the R translated equivalents, accompanied by GPT-4's explanatory comments.

Preliminary Results

Our research team ran a number of tests using a wide range of SAS programs to thoroughly evaluate GPT-4's capacity to convert SAS code into R. These programs included both basic components and more intricate structures.

Proc Steps: These procedures are core to SAS, dictating how specific tasks or operations are executed. It's paramount that the essence and functionality of proc steps are accurately captured in R.

Result: GPT-4 demonstrated a proficient understanding of various proc steps, delivering R translations that mirrored the original SAS procedures' objectives. Some complex steps, however, required subtle refinements post-translation.

Proc SQL: This provides a means to execute SQL-like commands in SAS. Given its uniqueness, replicating its intricacies in R is a challenge.

Result: GPT-4 managed to transcribe the majority of Proc SQL commands into corresponding R functions with a high degree of accuracy. Some advanced SQL commands necessitated minor manual tweaking to function seamlessly in R.

Macros: Representing the formalised structure of SAS, macros can be complex, encapsulating varied functionalities.

Result: The translations of SAS macros into R functions were predominantly precise, maintaining the intent and functionality. For some macros, especially those heavily reliant on SAS-specific utilities, post-translation enhancements were beneficial.

In conclusion, the utilisation of GPT-4 has demonstrated its significant value in facilitating the transfer from SAS to R. Although there may be certain elements that could potentially benefit from human involvement and improvement, the vast majority of the translations exhibit notable correctness and operational integrity. This facilitates the transition process, resulting in improved efficiency and effectiveness.

IT and HR Requirements

Development and IT Resources

- **RStudio:** Selected for its comprehensive R tools, it facilitated the application's scripting, testing, and debugging.
- **RStudio Connect:** Made it easier to deploy the shiny "SAS to R Code Assistant" to the local network, ensuring easy access.
- **Enterprise OpenAI Account:** Essential for harnessing GPT-4's translation capabilities, offering enhanced access, support, and security during interactions with the API.

HR requirement

- **Lead Data Scientist:** Played a key role by dedicating 25% of their time, ensuring project alignment with best practices and objectives.
- **PhD student**: a full-time participant who contributes substantial research and a structured methodology early on in the project.
- **Intern:** A full-time assistant who supports research, pilot experiments, and office work so that senior staff members can focus on the most pressing issues.

Barriers

The CSO required a business case to be presented to support the use of the OpenAI API. For the first use case, a clear strategy was laid out to make sure that only SAS code was provided to the application. The development and use of the application were restricted to a single section.

Stakeholders, Validation and Quality Metric

The "SAS to R Code Assistant" application has currently been made available to a test group within our office setting. The initial roll-out attempts to comprehend its practical applicability, pinpoint areas that should be improved, and gauge overall usability.

Feedback is actively sought from the pilot users, and efforts are now being made in this direction. Early signs point to a favourable welcome, with no significant issues raised. The program will be further improved thanks to this feedback, assuring its effectiveness and usability.

Important Evaluation Metrics

The application's capacity to accurately convert SAS code into R serves as the major metric of its effectiveness. Two crucial metrics serve as our assessment's guiding principles:

- The translated R code must successfully compile and execute without issues.

- It is crucial that the outputs produced by the R code match those from the SAS. The translation is accurate both syntactically and semantically thanks to this uniformity.
- To verify the accuracy of the results, the entire code base will be executed concurrently using SAS and R equivalents.

Collaboration

The "SAS to R Code Assistant" scope will be expanded upon when the code is revisited for collaboration. In response to requests from other governmental organisations or groups, we are prepared to open-source the application code.

Current stage

The "SAS to R Code Assistant" has been made available to a limited number of users in production and is currently in the pilot stage.

Other Potential Use Cases

There are numerous use cases that extend beyond code translation from SAS to R.

- **Code translation:** Facilitating translation between any two programming languages.
- **Standardisation:** Paving the way for unprecedented standardisation across various platforms and languages.
- **Code simplification:** Reducing and simplifying code complexity for easier comprehension and maintenance.
- **Enhanced explanations:** Providing clearer and more detailed explanations of translated code sections.
- **Efficiency gains:** Offering significant efficiency improvements, especially when handling and analysing big data.
- **Language to code creation:** Translating verbal instructions directly into executable code.

There are numerous potential use cases, but within the scope of this document, the focus has been narrowed to programming code exclusively.

Next steps

All pilot users of the "SAS to R Code Assistant" will be actively asked for detailed feedback. We intend to improve and polish the application based on their recommendations and insights to make it even more user-friendly. Optimisation is a priority, with the goal of enhancing usability. After these upgrades, we plan to broaden the distribution and make the tool accessible to more people inside the organisation.

3.3. StatGPT (IMF)

Business Problem and Value Addition

Over the last 25 years National and International Statistical Offices have transitioned their dissemination practices from print publication of statistics to digital creating significant benefits to users – both in terms of cost and accessibility. However, challenges remain including the practice of disseminating data in thematic silos. While users can readily access data for themes such as employment, turnover, prices, and output, combining data across themes is more problematic. For example, if a user requires labour data for the automotive industry it is relatively straight-forward to retrieve but if the user requires labour, output, export, and price data for the automotive industry it is more cumbersome. In most cases they would need to retrieve the individual series (conducting several searches) and then combine the data (usually off platform) to undertake their desired analysis. A second challenge users often face is the ability to undertake bespoke aggregations and calculations. For example, a user coming to the International Monetary Fund (IMF) website to examine Consumer Price Index (CPI) data may be interested in comparing the aggregate CPI for Canada, the US and Mexico with the CPI from the European Union (EU). If the IMF has not produced a Canada, US, Mexico aggregation the user would be required to compute this aggregation on their own – potentially using a different method than what is used for the EU total, resulting in an inconsistent comparison.

Generative AI has the potential to address both issues - enhancing user experience and the quality of their analysis. The IMF is currently modernising its statistics processing and dissemination platforms bringing these systems in line with modern best practices including faceted search and browse capabilities. When ChatGPT was released, and the power of generative AI became apparent, the team overseeing the modernisation effort began investigating if this technology could enhance the search capabilities of the platform. A proof-of-concept prototype (referred to as StatGPT) was developed that leverages Generative AI to assist users in accessing data on the dissemination platform. StatGPT uses a natural language interface, where end users communicate their request for data using plain English. The main task of StatGPT is to properly decipher and extract all the necessary parameters from a natural language prompt, use this information to construct query parameters and perform a data query request against an API that returns statistical data. It is possible that an end user's request for data could be misinterpreted (i.e. the final parameters for a data query do not properly represent user intent). To avoid this and increase the level of transparency, StatGPT displays the query parameters in plain English to the end user allowing the user to confirm the query parameters. In addition to permitting users to design complex multi-thematic queries and perform bespoke calculations, a Generative AI enabled search has several secondary benefits. First, the time to discovery should be greatly reduced as moving search from keyword to natural language should allow users to become familiar with the data holdings on the platform much quicker. Second, the StatGPT solution enables the collection of anonymized usage statistics and prompts that deliver valuable insight into how users interact with the data holdings. This information can be used to improve how the data are structured and ascertain which products are in demand (or not) and where metadata improvements are needed. The improved metadata will in turn help resolve ambiguity that is often present in a natural language request.

Description of Solution

The StatGPT project is divided into two phases. Phase one is focused on delivery of most foundational and user-relevant features weighted by the degree of their implementation complexity. Phase two (if phase one proves successful) will focus on integration with Excel (a popular tool used by IMF economists to integrate data into forecast models) and supporting simple data computations based on natural language commands.

The functional requirements for phase 1 include:

- Enabling end users to find and access IMF statistical data using a natural language. interface that constructs a SDMX data query.
- Assisting end users in finding the best-fit indicators based on user intent.
- Avoiding data misrepresentation by permitting end users to review query parameters prior retrieving data.
- Allowing users to query and combine indicators across several thematic datasets.
- Ability to visualise returned statistical data in tables and charts.
- Ability to generate and visualise SDMX REST API query and Python code snippets.
- Ability to use business glossary groups (e.g., G7, Advanced Economies, Oil-producing countries codelists) in end-user natural language requests.
- Integration of StatGPT Chat Interface with the IMF Data Dissemination Platform.
- Limit user intent to finding statistical data.
- Respect and enforce entitlements for data access.

Non-functional requirements for StatGPT:

- End users must be authenticated to use; anonymous users are not supported.
- Log and provide insights on end user query execution.
- Log and provide metrics on costs for utilising Azure Open AI API.
- StatGPT is deployed in Azure Cloud using Azure Open AI.
- Utilise iData SSO for end user authentication.
- Load balancing and resiliency.
- Mobile friendly (responsive design).
- Per user cost throttling (required user authentication).

StatGPT solution consists of front-end and back-end components communicating via REST APIs. Front-end component (StatGPT Chat Interface) provides user interface (UI) for end user input as well as output that can embed visual elements such as tables, charts and code snippets. This component can be embedded in a given dissemination platform (i.e the IMF Data Dissemination Platform or IMF Data Excel Add-in).

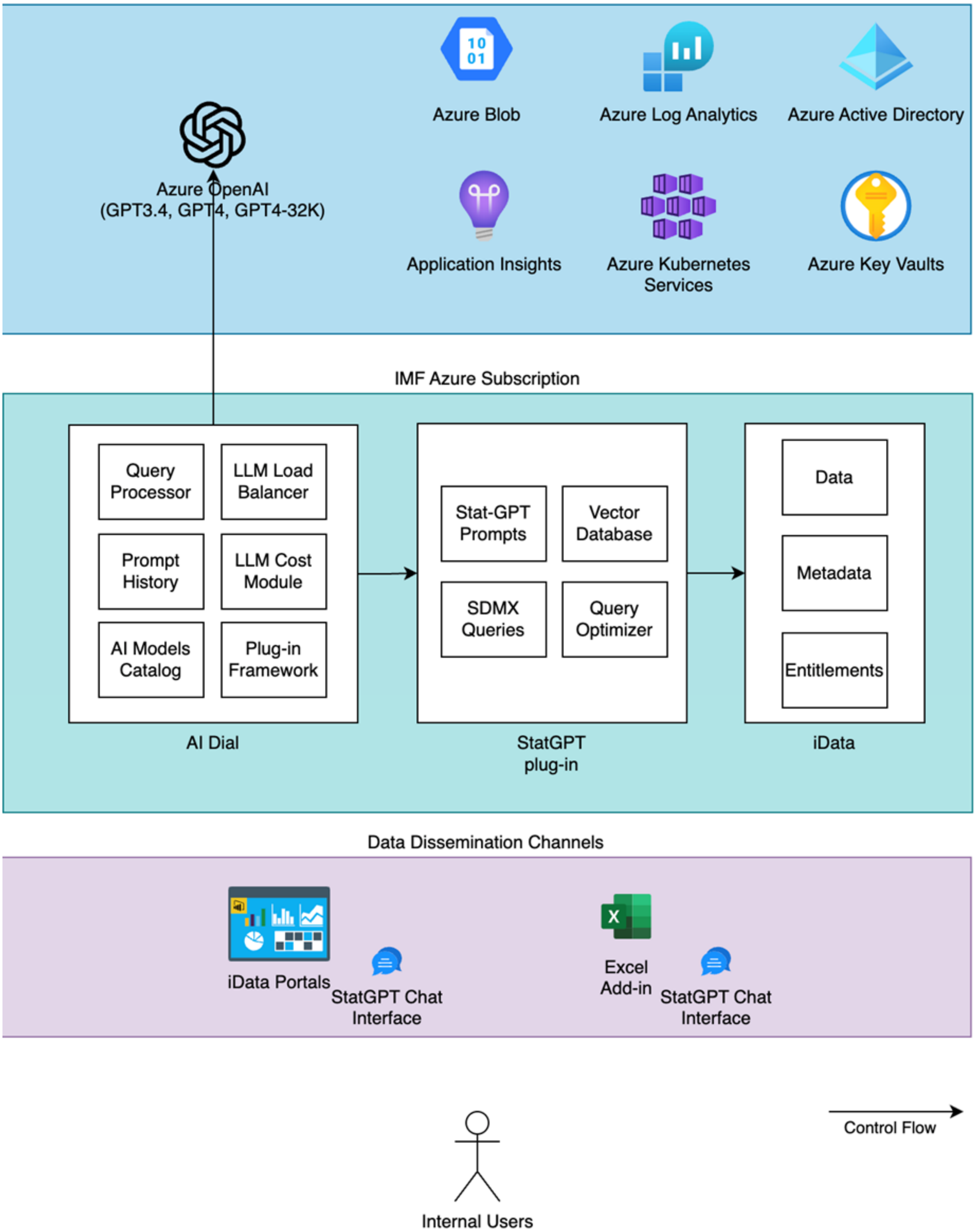
Back-end components consist of a general AI-Dial component (developed by an external IT vendor, EPAM) and IMF specific StatGPT plug-in component. The AI-Dial (https://epam-rail.com/) (Deterministic Integrator of Application and LLMs) component is responsible for the operational aspects of the solution like query load balancing, LLM model selection (including GPT3.5/GPT4/GPT-32K), usage logging and analysis. Choice of a specific LLM model to be used in StatGPT is determined based on cost vs features trade off analysis to be conducted during Phase 1. The main goal of the back-end components is to generate a proper prompt for the selected LLM model that contains relevant IMF context and proper instructions to answer end user questions. Prompt instructions specify three parameters needed from a user dialog in order to complete a query definition (indicator list, country list and time period) and these lists need to be shown to an end user. IMF context consists of:

- all indicators and their metadata, including their dataset metadata, from all of the datasets enabled.
- all countries and their metadata, including country groups and their breakdown.

This context gets vectorised using LLM embedding service and stored in a vector database. To answer a user question, back-end components perform the following execution flow:

- extract all the context from a vector database that "matches" a user question, which ultimately results in a list of indicators and countries along with their system IDs.
- generate a prompt that includes a user question, prompt instructions, extracted context and additional instruction for LLM to use extracted context for its response.

This execution flow repeats until all query parameters are defined, provided by LLM response completion, at which point a user can execute a constructed query and get the data. To make query completion easier, the LLM can be instructed to use a specific time period as a default so that end users can omit specifying this query parameter in their question.



Preliminary Results

The IMF, in conjunction with a private sector vendor, EPAM, has developed a prototype application that allows end users to query IMF datasets via the existing SDMX API using natural language. The prototype was limited to querying a single dataset – the IMF's World Economic Outlook (WEO). While the success rate of the tool is extremely high in certain cases, the query parameters did not always yield the expected results. For example, assume the end user requested StatGPT to provide the "All items CPI for French speaking countries."

StatGPT was able to identify a list of countries but in some cases StatGPT included an English-speaking country and in other cases missed French-speaking countries. Similar issues arose with requests for bespoke industrial or commodity groupings. The next iteration of the project will attempt to address these issues by providing the Generative AI information related to geographies, industries, commodities, statistical methods, and methodologies that should limit the "AI knowledge gap" and increase the accuracy of the query parameters.

IT and HR Requirement

The proof of concept and phase 1 is being developed by IMF IT and business experts and an external IT vendor. The initial prototype for the proof of concept was built by a small team in a short time period (months). Phase 1 will consist of a team of 8-9 resources working over a period of 3-6 months.

Validation, Human in Loop and Quality Metrics

To ensure users are provided with the data that match their request end users are asked to validate the query parameters along various dimensions (e.g., country, industry, indicator, time) prior to submitting the query and retrieving the data. In some cases end users also use StatGPT to request clarification of terms, definitions of variables or to understand the sources of methods for a given indicator. For these types of requests the Generative AI will respond directly to the users' request but there is a possibility that the answer will be imprecise or incorrect. To reduce the possibility for imprecision, in phase 1 of the project the development team will ensure StatGPT does not provide a response to these requests either by employing the GPT "system role" instruction within a prompt (e.g., You are an economist working at the IMF and your mission is to make sure all questions are only related to macro economics) or placing small language model in front of user requests that decides, based on training, what is an "allowable" question to be passed to the large language model. For phase 2 of the project the development team is exploring the possibility of limiting the Generative AI's response to a set of selected methodological documents such as the Balance of Payments Manual, the System of National Accounts etc. that contain the concepts, classifications, definitions and sources and methods that underpin the datasets on the IMF's dissemination platform. In addition to limiting the response to economic and environmental accounting manuals, phase two will also explore the possibility of leveraging the metadata associated with the specific datasets. It may be possible to have StatGPT develop query parameters for the IMF's metadata holdings when the end users' prompt is related to concepts, classifications or sources and methods. For example, assume an end user requests a definition for "GDP" and that there are three datasets on the platform that contain an estimate of GDP, each with a slightly different definition. Prior to providing the definition, StatGPT could ask the user if they want the definition from dataset 1, dataset 2 or dataset 3 (of course the fact that there are three definitions of GDP may indicate a data governance issue on the platform but at least StatGPT would not be making up a fourth).

Stakeholders

StatGPT has been shared with IMF internal working groups and presented at several international conferences. Most stakeholders agree there is a strong business case for leveraging Generative AI to query statistical databases and that it will result in improved access and increase the use of statistical information.

Barriers

Given this work involves the integration of a new technology into the IMF, the development of the product must align with the development and evolution of corporate policies around Generative AI. As such, the development of StatGPT is somewhat slower than would be the case if the underlying technology was already well established within the organisation. In addition, the project is tied to a larger modernisation project impacting the availability of resources and overall timeline.

Risks and Mitigation

The key risk associated with this work is the risk of providing erroneous data to end users. This risk has been mitigated in two ways. First, by using the Generative AI to construct query parameters and passing those parameters to a structured dataset ensures accurate responses. Second, having the end users review the query parameters (in natural language) before obtaining the data ensures the user are getting "exactly what they are asking for". There is also a risk that users will ask StatGPT to respond to "non-queryable" questions such as "how is GDP calculated." This risk has been reduced for phase one by limiting the ability of StatGPT to only respond to data queries and providing a response of "I cannot respond to your request at this time" when prompted with a non-data retrieval question. To reduce exposure to reputational risk associated with poor responses any application considered for release to the public would first undergo significant testing by internal users.

Collaboration

This work is being undertaken by the IMF's Statistics Department, IT Department and an external vendor, EPAM, that is developing the IMF's data processing and dissemination platforms. The IMF hopes to be able to share this work with the greater SDMX community as a start, but the application could be implemented to query any structured dataset accessible via an API.

Current Stage

An enhanced proof of concept is being developed with the features outlined above. The IMF is currently working with the vendor to stand up a component that can be added to its internal dissemination platform that is expected to be launched in the spring/summer of 2024.

Other Potential Use Cases

As currently envisioned StatGPT will only query datasets within the IMF dissemination environment. In the future it is conceivable that someone that comes to the IMF dissemination platform could use StatGPT to query data on other SDMX platforms such as the World Bank, OECD, or Eurostat (or National Statistical Offices).

Next steps

The next step is to complete phase 1 development and begin testing with internal users. The IMF is also proposing to present this at the upcoming SDMX Global Conference in the fall of 2023. If phase one proves successful the IMF will launch phase 2 of the project which will add additional features such as 'on the fly' bespoke calculations.

3.4. Report Generation Using LLMs (Statistics Canada)

Business Problem and Value Addition

Effectively conveying complex data to users through informative articles, data tables, and images is essential for facilitating understanding, enabling evidence-based decision-making, and unlocking the full potential of our data products. However, the current process of creating these descriptions is resource-intensive and dominated by repetitive tasks. This not only leads to potential delays but also diverts our analysts from more in-depth, valuable work they could be performing.

To address these challenges, we are exploring advanced machine learning techniques, specifically text generation, to develop an innovative data-to-text generation model. The main goal is to efficiently generate text descriptions for Census data tables, enhancing their interpretability and improving their timeliness

It's imperative to emphasise that this process serves to create an initial draft of a report. The generated content will undergo review, vetting, and editing by a human analyst before finalisation. A human-in-the-loop validation is an integral part of this process, ensuring the highest quality of output.

Description of Solution

Our LLM-based solution is designed to cater to various data types and user requirements. It goes beyond the mere description of data points, providing capabilities to highlight trends in variables over time and perform comparative analyses between variables such as provinces, age groups, etc. Drawing upon the extensive pretraining of the model on diverse text sources, it can generate articles across a wide range of indicators. Carefully crafted prompts guide the model to produce articles in the style consistent with Statistics Canada's previous publications.

We explore and evaluate various innovative data-to-text pipelines powered by LLMs. The objective is not only to uncover the strengths of these approaches but also to identify and address any potential limitations or biases in the generated text. This thorough evaluation provides a profound understanding of the capabilities and constraints of text generation methods within the unique context of Statistics Canada releases.

The insights gained from these evaluations serve as essential resources for shaping future decisions regarding the adoption of text generation technology. Additionally, they contribute to the development of expertise within Statistics Canada, empowering the organisation to fully harness the transformative potential of this technology.

Preliminary Results

One of the prominent challenges associated with LLMs and generative AI is the risk of hallucinations, where the model generates invented data or facts to enhance its responses. This poses a significant concern, especially in the context of creating statistical articles, where factual accuracy is paramount. To mitigate this risk, our approach emphasises grounding the model in verifiable facts.

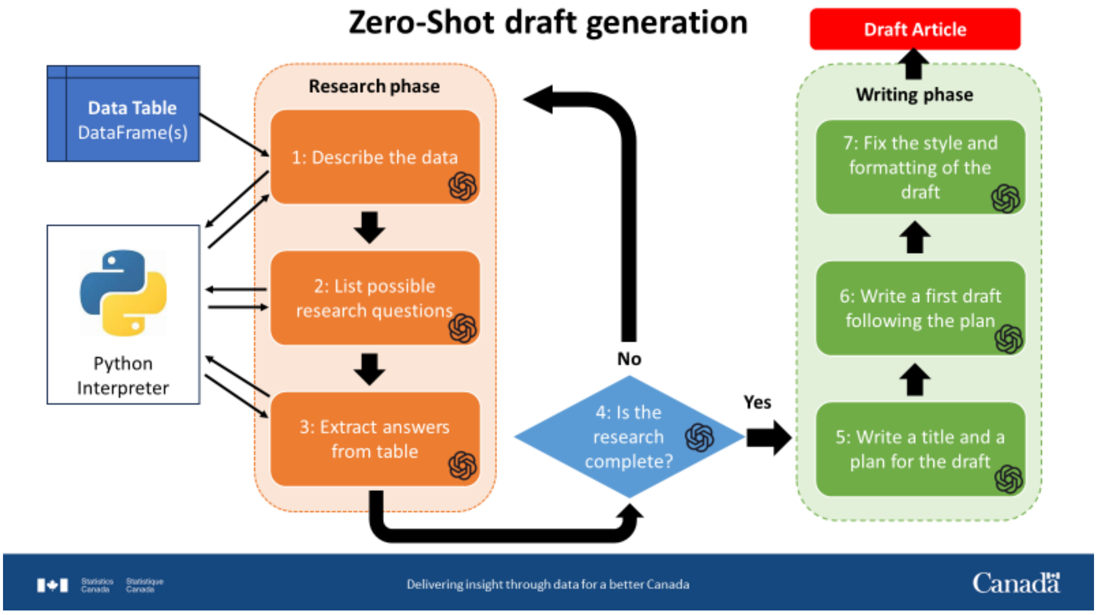
Rather than directly feeding raw data tables to the model, we provide it with table descriptions and headers. We then task the model with generating the necessary Python code to extract facts directly from the tables. This code is executed locally on the same machine where the data resides. This meticulous process ensures that the generated content is firmly rooted in actual data, eliminating any reliance on what the model may remember from its training data.

The architectural design of our solution follows a systematic step-by-step approach, divided into a research phase and a writing phase. Multiple quality verifications are integrated, performed by the LLM itself, to ensure a robust and predictable output.

In the research phase, the model is first prompted to describe the data, leveraging Python code to extract summary statistics from the tables. It then generates a list of research questions, for each of which it creates and runs a Python script locally, verifying if the output effectively answers the question. Once all research questions are addressed, the model generates a comprehensive research report that serves as the foundation for the model to produce the article.

The article-writing phase employs a similar approach, utilising inputs from the research report and drawing inspiration from past articles for writing style. It initiates by generating a detailed plan and then iteratively produces drafts until it achieves the final article.

The initial results obtained using GPT 3.5 Turbo are promising. The generative pipeline successfully produces factual articles deeply rooted in the provided tables. However, ongoing work is essential to enhance the style and coherence of the articles. Notably, there is a need to address issues such as excessive article length and a writing style that could benefit from a more concise and journalistic approach.



IT and HR Requirement

This project involved a team of three data scientists working over the course of four months in a Microsoft Azure cloud environment. They accessed GPT-3.5 and GPT-4 models through the Azure OpenAI Service. Billing for this service is consumption based, and total cost of using the OpenAI models was a few hundred dollars.

In addition to cloud resources, access to Graphics Processing Units (GPUs) was a pivotal requirement. These GPUs were essential for the exploration of open-source LLMs, including Meta's Llama 2. Provisioning GPUs on cloud platforms can be a costly endeavour. However, to optimise resource usage and minimise costs, Statistics Canada leveraged existing GPU servers located on-premises, streamlining the computational needs of the project. This strategic use of existing infrastructure contributed to cost-efficiency.

Validation, Human in Loop and Quality Metrics

For robust validation and quality assurance, we employ the Arthur AI Bench, an open-source evaluation tool specifically designed for comparing various LLMs, different prompts, and hyperparameters. This tool offers multiple scoring options for evaluating general generative tasks, encompassing Prompt-Based Scorers, Embedding-Based Scorers, and Lexicon-Based Scorers.

During testing phases, our generated text undergoes rigorous evaluation, including comparisons with gold standard text using BERT Score for similarity. Additionally, we assess the readability and specificity of the generated content using Lexicon-Based Scorers.

To ensure a comprehensive assessment, manual validation is an integral part of our process. This critical step involves active participation from Census Operations and Subject Matter Analysts. They provide specific feedback on various aspects, including content quality, writing style and structure, factual accuracy, analysis, observations, and the overall flow of the content. This human-in-the-loop validation process enables us to gain a nuanced understanding of each subject matter focus area, allowing us to fine-tune prompts and continually improve the quality of our generated content.

Stakeholders

This project is supported by the census dissemination project composed of members from census operations and various Subject Matter Areas (SMAs) across Statistics Canada. The endeavour to explore generative AI tools, aimed at assisting SMAs in swiftly producing concise content that includes data tables without requiring in-depth analysis, has garnered widespread interest among stakeholders.

The initial response from Census Operations has been notably enthusiastic and positive. Similarly, SMAs have expressed keen interest in experimenting with the utilisation of such a tool in their standard processes. Stakeholders have various viewpoints on how to best utilise such a tool.

Our goal is to provide conclusive evidence on the best use case through a series of experiments and validation processes. This approach ensures that we align the capabilities of generative AI with the precise needs and preferences of our stakeholders, thereby maximising its utility within Statistics Canada.

Barriers

Statistics Canada has one of the largest cloud workloads among all Government of Canada departments. Consequently, the introduction of new technologies into the cloud infrastructure raises paramount concerns regarding cybersecurity and potential impacts on existing architectural and operational processes.

Gaining access to a new cloud service in Statistics Canada's cloud requires a review by the Cybersecurity Division and discussions with the appropriate committees and boards responsible for enterprise architecture. This process is still ongoing for access to the Azure OpenAI Service and is a major organisational challenge for innovative projects looking to employ new technologies.

To circumvent these barriers, our project team opted to work within a separate cloud environment external to Statistics Canada. This alternative cloud environment proved to be both accommodating and conducive to supporting our project's objectives. It is important to note that, especially given the external cloud environment, the data used for this project was all previously released and not protected or confidential.

Risks and Mitigation

One significant risk inherent in our project revolves around the fact that the current cloud environment in use does not belong to Statistics Canada, leaving us with limited authority and control. To address this concern, our IT team is diligently working on the development of a parallel cloud environment within our own tenant. This tailored environment will provide us with the necessary access to Microsoft Azure Open AI while ensuring that we maintain a high degree of autonomy and oversight.

It's important to note that the project will not be able to transition into a production phase for the use of pre-release data until this environment is fully established within our domain. Moreover, we must await approval from the Canadian Centre for Cybersecurity, validating the compatibility of Azure Open AI with the handling of protected and confidential pre-release data. These measures are integral to the comprehensive risk mitigation strategy, ensuring the security and compliance of our project as we progress toward its ultimate objectives.

Collaboration

Statistics Canada hosts an interdepartmental Applied Text Analytics and Generative AI Community of Practice (CoP) to facilitate collaboration and knowledge sharing across the government. This CoP hosts over thirty agencies and departments and hundreds of active participants monthly.

The CoP is among the leading forums for discussion and collaboration around generative AI in the Government of Canada. Recent presentations have prominently featured projects involving generative AI from a variety of domains, including justice, agriculture, and defence.

Notably, at least twelve departments are actively engaged in exploring the potential of generative AI.

Furthermore, our project benefits from a strategic collaboration with Shared Services Canada (SSC), the centralised provider of information technology services for the Government of Canada. SSC hosts the Science Program, a dedicated cloud platform tailored for exploratory scientific work, available to Science-Based Departments and Agencies across the government. This program grants access to an innovation cloud environment that operates independently from any other government infrastructure, allowing for lighter administrative and security requirements. As part of this collaboration, SSC has provided Statistics Canada with its own Azure subscription within this environment which enabled Statistics Canada to use OpenAI models.

Current Stage

At present, we are actively engaged in an experimental phase, wherein we are rigorously exploring the capabilities of various LLMs. This includes OpenAI's GPT 3.5 and GPT 4.0, alongside open-source LLMs like Llama2.

Alongside our experiments, we are diligently tuning our prompts. This process is driven by the invaluable feedback we receive from our client areas, which helps us in crafting relevant articles.

Other Potential Use Case

While our current focus revolves around census data and articles, the versatile capabilities of LLMs open the door to broader applications. These LLMs hold the potential to generate articles not only for census data but also for various other data sources within Statistics Canada. This prospect paves the way for extending the benefits of text generation to a wide array of domains, enhancing the utility of LLMs across our organisation.

It's essential to differentiate between the types of content these models can produce. LLMs can efficiently generate summaries from diverse data sources within Statistics Canada, condensing information for easier comprehension and review. Additionally, with the right training, fine tuning and context, LLMs have the potential to generate deeper insights, though this requires careful consideration and validation, and is out of scope of our work at this stage.

Next Steps

Close collaboration with our client areas and working in an iterative manner is essential for the success of this work. We are currently collecting feedback from numerous areas to adapt and adjust the process, prompting and output layouts. Additionally, we have started working on creation of alt-text for images and graphs that may exist in some reports. Creation of simple descriptive alt-text which can be verified by an analyst quickly will be essential for accessibility purposes.

Furthermore, we are exploring the integration of LLMs for translating texts within the generated reports and alt-text. This ensures that our initial drafts will be readily available for review in both official languages, English and French.

In parallel, we are closely collaborating with our IT colleagues to establish a secure cloud environment conducive to our experimentation. This includes working within our internal cloud environments, where access to Microsoft Azure Open AI and open-source LLMs is facilitated. Notably, the current cloud environment we are utilising is external to Statistics Canada and is not suitable for scaling up or transitioning into a production phase.

Moreover, we remain committed to upholding responsible ML (Machine Learning) practices and ethical standards, guided by Statistics Canada’s Responsible ML Framework and the principles outlined by Treasury Board Secretariat. Our commitment to ethical AI is further underscored by adhering to the guidelines outlined in the Government of Canada's Guide on the Responsible Use of Generative AI, through ongoing collaborations with the Treasury Board Secretariat as well as preparation of Statistics Canada’s own guidelines on use of Generative AI which are currently being drafted. These considerations underscore our dedication to ensuring the integrity and ethical soundness of our project as we advance into the next phases of development and implementation.

3.5. Metadata Editing Leveraging GPT (BIS)[11]

Business Problem and Value Addition

The Bank for International Settlements (BIS) regularly disseminates statistics through the BIS Data Portal on major financial and macroeconomic indicators, spanning from credit to

[11] The views expressed are those of the authors and do not necessarily reflect those of the Bank for International Settlements. All errors are our own.

the non-financial sector, property prices, exchange rates and international banking statistics. Data are released along with extensive metadata attached to observations, time series or a data set. The dissemination of these metadata, including information about methodology, collection, coverage and sources, is instrumental to promote and enhance the understanding of the BIS statistics. Leveraging the ISO standard for SDMX, metadata are embedded in a consistent, orchestrated and homogenous way across several features in the BIS Data Portal, such as dashboards, tables and the glossary. This approach prevents content duplication and helps users to navigate through complex information quickly.

Given the pivotal role played by metadata, their quality management is crucial. However, it is well known that metadata editing is an extremely time-consuming task, as it often requires manual review by statisticians. This process typically involves several checks, spanning from the application of basic rules, such as capitalisation, formatting and layout, to more advanced tasks such as grammar, spelling and syntax checks, verification of the fluency and consistency across the text and review of the overall logical flow. Depending on the complexity of the text to be checked, a statistician may spend several minutes to validate the value for each attribute per time series. We estimate that, on average, the full editing of one attribute with 200/250 characters takes between 30 seconds and 1 minute per statistician.

To cope with these challenges, often arising from limited capacity, the BIS DataBank team – which oversees a number of data sets published on the BIS Data Portal - is actively exploring innovative AI-driven solutions to streamline the metadata editing process.

Description of Solution

The BIS DataBank team is currently focusing on developing AI-powered assistants that respond to specific sets of instructions to edit the metadata fields. More specifically, the metadata editing process is orchestrated in several key steps.

Firstly, we initiate the procedure by specifying our requirements. For instance, we ask for applying English capitalisation rules. In addition, we also indicate specific rules, such as to ensure the precise spelling of names associated with central banks and other institutions, following the official names of the BIS shareholders. This is a crucial step to eliminate any potential error that may compromise the integrity of the metadata.

Secondly, the process also involves a comprehensive formatting clean-up. This includes the removal of double, trailing, or any other extraneous characters, thereby ensuring a standardised and polished text. While this step may also be achieved using regular expressions, the added-value of leveraging an AI-assistant is the ability to preserve metadata consistency thanks to a throughout examination of the text. This involves not only cross-referencing *within* but also *across* attributes of time series to ensure uniformity. For instance, we apply consistently "data are sourced" instead of "the sources of these data are". Other examples may involve applying consistently date formats or replacing abbreviations.

Thirdly, we also give instructions to check for grammar, typos and other syntax errors. The instructions may also include other verifications, such as shortening the text in case it exceeds the counter limits – which are usually explicitly set as part of the SDMX data modelling and could be an issue for statisticians while compiling metadata.

By implementing these steps, our metadata refinement process aims at delivering better readability, visual clarity and overall coherence. We apply these instructions to different types of metadata attributes, such as titles, compilation, coverage and source which are usually released to the public domain.

Preliminary Results

At this stage of the project, which is still in its infancy, we can report promising results that carry significant potential, especially to boost productivity. We tested our alpha AI-assistant, both on GPT 3.5 turbo and GPT 4.0, on two data sets. The first one contains around 300 time series whose key attributes (title, coverage, collection and source) require significant editing (e.g., wrong capitalisation, typos and grammar checks, low quality phrasing). The second one contains around 400 time series whose metadata are assumed to be error-free but required phrasing / consistency checks. Most of the time series included in this data set feature free text attributes exceeding 500 characters (with a maximum set to 1050), thus involving substantial work. We took advantage of the OpenAI API to programmatically return the output generated by the assistant against our prompts while the system instructions are embedded into the assistant.

Our preliminary findings indicate that this solution is generally carrying the expected results. Basic tasks, such as capitalisation, punctuation, whitespaces or other extraneous character are easily detected and correctly processed. The formatting of dates also features excellent results. We note, however, some incorrect outputs when it comes to the handling of plural/singular forms (for example, when explicitly instructing to read “data are…”, the assistant also erroneously changed to other plural words whose form shall be singular in the given context such as “series”). Furthermore, the handling of specific words, including the official denomination of central banks, sounds very promising, although requires more fine-tuning with appropriate training.

IT and HR Requirements

HR Requirements

The design and implementation of this project required, to date, a small team made of one statistician and one associate statistician. Around 30% of the time was spent to assess and design the process, 40% on the coding and implementation and 30% on testing/benchmarking against other solutions.

IT Implementation

The IT implementation required an access to the OpenAI Enterprise account and evolved with the maturity of the project. The use of playgrounds was the most efficient approach during the pre-assessment phase as it enabled a quick calibration of the AI-assistant without high barriers and fixed implementation costs. The use of a custom ChatGPT was also part of the exploration but quickly reached its limitations (e.g., tokens, modularity). Once the pre-assessment was successful, the team moved to more advanced tools to tailor the app to the user requirements, for example, involving Python scripts with more advanced techniques (asynchronous calls, etc.).

Barriers

In the use case described in this section, only public metadata are exposed to the OpenAI API. However, similarly to other official organisations, the most critical barrier for the further expansion of the envisaged solution is the confidentiality restrictions that may apply to the use of the API. Other barriers include the limitations to the number of tokens to be passed per request / minute as well as the lack of integration of fine-tuned models with assistants in the OpenAI, which is not offered at the time of writing this document.

Stakeholders

The proposed solution holds particular significance for two key stakeholders: statisticians and IT developers. On the one hand, statisticians will benefit significantly from the streamlined processes, as these advancements are poised to save them valuable time currently taken up by repetitive and highly manual tasks. On the other hand, IT developers may find these developments instrumental in advancing their innovative and efficient solutions, for example, in the context of building new metadata editing and validation pipelines. The synergy between these two stakeholders is critical to make concrete advances in this domain.

Validation and Quality Metrics

This solution features two distinct types of validation rules to ensure the accuracy and integrity of the returned metadata against the instructions passed to the system role.

The first one may contain objective validation rules, for example, involving automated checks designed to verify specific criteria. Examples include validating whether returned sentences end with a period or ensuring that any word following a period begins with a capital letter. These rules are easily measurable, providing an efficient way of enforcing standardised formatting. To date, all the returned results passed this first chunk of validation rules.

The second type of validation rules is subjective, for example, to ensure that the meaning and/or context of the metadata remained unchanged. This validation step requires a review by a statistician. To facilitate the process, we leverage metadata version-control that is natively embedded in the BIS DataBank software to review and audit input metadata submissions.

Risks and Mitigation

The core operational risks mostly relate to the possibility that the solution may not work as intended or cannot be accessed, leading to disruptions in the workflow. To mitigate this risk, we are further testing our solution. We also note that, to date, this solution only applies to one-off exercises as metadata editing usually occurs occasionally, not on a stream basis.

The proposed solution also carries reputational risks, especially associated with the risk of disseminating inaccurate metadata. To mitigate this risk, we keep humans in the loop. To ensure the quality of the metadata, statisticians review the final version of the metadata, and we also leverage version control and audit trails to spot potential mistakes.

Other Potential Use Cases

Building upon the promising results obtained so far, it might be conceivable to extend our solution also into metadata validation, beyond editing. One such advancement may involve harnessing the SDMX logic to check, for example, the coherence between free text attributes and enumerated codes in the series code (e.g., to check whether a series with the code "5B0"[12] for the dimension source matches the free text "Bank for International Settlements" in another related attribute).

Next Steps

The next immediate steps involve conducting further testing, creating a streaming application and applying this method to a wider sample of data sets. These next steps will be key to refine and validate the effectiveness of the method across a broader spectrum of cases, while also harvesting more advanced technical solutions (e.g., LangChain).

In the medium run, assuming that the method proves to be robust enough and the identified risks are successfully mitigated, we may augment this method into a systematic and automated workflow for handling metadata, along with other data-related processes.

[12] https://registry.sdmx.org/sdmx/v2/structure/codelist/IMF/CL_ORGANISATION/1.13

RISKS AND MITIGATION MEASURES

04

4. Risks and Mitigation Measures

In addition to the opportunities that LLMs offer, there are also risks that need to be taken into consideration. In this section we describe risks arising when implementing LLMs in statistical organisations and potential mitigation measures. Starting with ethical considerations, we explore accuracy issues, privacy and security concerns, legal complexities including copyright issues, and potential misuse arising from lack of literacy.

4.1. Ethics

Ethical obligations extend beyond legal obligations in all aspects of life. For statistical organisations that produce official statistics, maintaining high ethical standards is particularly important as it reinforces public trust and social licence to operate. There are multiple relevant ethical frameworks for the use of AI[13] [14] [15], all of which cover similar principles. The Principles for the Ethical Use of AI in the United Nations System was used as a basis for this analysis. Table 4.1 discusses the relevance of each ethics principle to the use of LLMs in the statistical organisations. Broadly, the major considerations therein are the protection of human rights and the need for human-centred oversight and authority.

Table 4.1: Relevance of Ethical Principles to the use of LLMs

Principle	Relevance to use of LLMs
Do no harm	Care should be taken to ensure that LLMs are never used for purposes that cause harm. This may be intended or unintended harm, such as unconscious biases built into LLM responses. To address this, projects involving LLMs should have defined expected benefits and a risk register with mitigations for any potential harms that may arise.
Defined purpose, necessity and proportionality	The use of LLMs should be proportionate to the needs of the organisation and should not overreach or undermine human authority or human rights.
Safety and security	Safety and security risks should be considered for all projects involving LLMs. This will include information security risks, but may also include cyber security, as advanced malware seeks to take advantage of LLM technologies.
Fairness and non-discrimination	As a function of their underlying training data, LLMs are subject to intentional and unintentional biases, which may be as simple as the language the training data is written in.

[13] https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf
[14] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[15] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework

	Care should be taken to evaluate and address biases in LLM outputs.
Sustainability	The impact of the use of LLMs on current and future generations should be considered as part of the defined benefits and risks of an LLM project. The environmental impact of training, running, and maintaining LLMs (e.g., the carbon footprint of the intensive energy consumption, water usage for cooling) and the implications for communities where the data centres required for these models reside should be especially considered.
Right to privacy, data protection and data governance	Rights to privacy, data protection and data governance apply to both LLM training data and the information supplied as prompts. Users should consider and evaluate the training sources used for LLMs to confirm they uphold these rights. Rigorous data governance rules should be maintained such that no sensitive information is released externally through the use of LLMs. This is discussed further in section 4.3.
Human autonomy and oversight	Decisions made by NSOs impact the freedom and autonomy of the citizens of their respective countries. as such, it is critical that human oversight, review and ultimate authority is maintained when using LLMs. No decision making ability should be ceded to LLMs. Under current usage conditions, a suitably qualified human should always be in the loop, and have a final say on any LLM generated output.
Transparency and explainability	Material produced through the use of LLMs should be disclosed in a manner such that no user could be mistaken as to its origin. Such a disclosure should also contain information as to how that material was produced and any limitations it has, including potential for bias, hallucinations, etc.
Responsibility and accountability	Similar to human autonomy and oversight, responsibility and accountability governance structures should reinforce the human-centred oversight and authority surrounding the use of LLMs and that those same people are responsible and accountable for any harms caused by the use of the LLM.
Inclusion and participation	All stakeholders must be consulted and considered surrounding the use of LLMs. This is particularly relevant in the context of a decision making process. LLM generated materials must only be one source of data, with humans provided an opportunity to contribute and influence over and above the AI.

4.2. Accuracy and Lack of Validation Mechanism

With any model, modelling system, or algorithm, determining and understanding accuracy involves a broad set of considerations. These are often dependent on the types of use cases the models are employed for.

For example, when considering a traditional machine learning model used in a binary classification problem, we may judge accuracy based on how well the predicted classifications match already known classification values. These can be quantified using a variety of measures, including confusion matrices, Brier scores and the like.

However, as the outputs, and importantly, the use cases demanded of a given class of models broaden, the question of accuracy also becomes more complex. The accuracy of large language models, particularly Generative Pre-trained Transformers (GPTs), therefore, becomes a complex and multifaceted topic. Assessing the accuracy of these models involves considering various dimensions, including their ability to generate coherent and contextually relevant outputs, their factual correctness, and their potential for biases.

Because biases are one of the most significant challenges with large language models like GPTs (and in various other types of machine learning in general) we explore this aspect of accuracy as its own section, given its importance to the work of statistical organisations.

Coherence and Contextual Relevance

The most striking driver of the ubiquity of GPT models appears to be their ability to generate text that is (or at least seems) contextually relevant and coherent. They are trained on massive amounts of text data, which allows them to understand and mimic human language patterns effectively. In general, they are highly accurate at generating text that appears contextually appropriate and natural. However, they may sometimes produce nonsensical or off-topic responses, especially when pushed outside their training data.

Factual Accuracy, Outdated Data, and Lack of Validation Mechanism

The factual accuracy of GPT models is a critical concern. While these models can generate text that sounds plausible, they may not always provide accurate information. In this context, information refers to the outputs from these models, be they paragraphs of text, code snippets, or other such outputs.

GPT models, including their larger iterations, do not have an inherent validation mechanism to verify the accuracy of the information they generate. They lack the ability to confirm or cross-reference facts or test code they are delivering for accurate performance. While there are efforts underway to improve their performance in terms of domain specific accuracy, it is difficult to ascertain their performance for specific domains or tasks ahead of time. Therefore, any use cases where these models are employed in a production pipeline requires formulating use case specific validation metrics. These are further discussed in Section 5 of this paper.

Furthermore, in general GPT models do not have real-time access to the internet and are not updated in real-time, which means that their “knowledge” is limited to what they were trained on. Additionally, they can sometimes generate information that is outdated or incorrect. Therefore, it is essential to fact-check information obtained from GPT models, especially for critical applications.

Numerical Capability and Accuracy

The numerical capabilities of LLMs, including GPTs, is a concern where accuracy and precision in data analysis and mathematical tasks is important. Because LLMs are mainly trained on textual data to understand patterns in language, they do not possess an inherent capability to deal with numerical data in a way that specialised algorithms and statistical/mathematical models do.

Additionally, because LLMs do not have an inherent capability to undertake complex mathematical operations, nor the training data to understand complex mathematical prompts (or complex numerical data), the usefulness and accuracy of processed outputs will be limited. For example, a summarisation of a data table that is deterministically accurate via a simple script in R may yield inaccurate, or in instances miss-predicted outputs when processed through an LLM. It should also be noted that LLMs do not have an internal framework of logical computation, therefore any derivation of mathematical functions or proofs are generated by maintaining patterns expressed in textual examples that the models have trained on.

Use-case Specific Accuracy Assessments as Mitigations

Use case specific assessment can mitigate the risks associated with these accuracy limitations discussed in the preceding sections, as demonstrated in Section 3 by various agencies.

For example, the ABS uses a modified BERT-Score to account for entire sentences in checking for semantic similarity for a use case for updating statistical classification definitions.

Statistics Canada's use case in Section 3.4, where some of the challenges related to these limitations are also discussed, uses human-in-the-loop type mitigations, coupled with various language testing protocols (e.g., BERT, Lexicon based readability assessments and the like) in a project attempting report generation using LLMs.

Model and Concept Drift

An additional consequence of outdated training data (apart from the inability to generate factually accurate or up to date information) is the risk of model and concept drift (which increases as the data becomes more out of date).

Model drift, where the performance of a machine learning model degrades over time due to changes in the underlying data distribution, is an ever present problem. In the context of large language models, when the patterns or relationships in the language change over time, models, which were trained on historical data, may become less effective.

For instance, if a language model is trained on text data from a specific period and is later used to analyse text from a different period, it might not perform as well due to the evolution of language, the emergence of new slang, or changes in writing styles. With the ubiquity of social media trends potentially driving fast changes in language patterns (e.g., new slang emerging) this may pose an increasing threat.

Concept drift refers to the situation where the relationship between input features and the target variable changes over time. For LLMs, this could mean that the meaning or context of certain words or phrases evolves. For example, the sentiment associated with a particular word may change over time, leading to a shift in the concept the model was initially trained to understand.

Data drift refers to the situation where the input data itself changes over time, which can lead to performance degradation. In traditional machine learning applications, in contrast to concept drift, data drift can be more readily monitored by maintaining a strong relationship between regular test data and production input data. However, with LLMs, this may prove more difficult due to the previously discussed lack of validation mechanism. This also introduces the risk of a degradation spiral, as the content balance that LLMs are trained on shift from predominantly human generated content to machine generated content, as machine generated content becomes ubiquitous.

Model drift, concept drift, and data drift are critical considerations in the deployment and maintenance of large language models, especially in dynamic environments where language use is subject to continuous change. The lack of inherent validation mechanisms in LLMs also exacerbates the already difficult problem of detecting these drifts and the associated degradation in accuracy and precision.

The risks associated with these are heavily dependent on the specific use cases. For example, in the case of legacy code conversion, such drift is unlikely to pose significant problems, given that the relational structures between a legacy language and a stable modern language are also likely stable.

However, where LLMs may be used for interaction with the general public, for example, where the public can use LLM assisted tools to interrogate official statistics more effectively, both model and concept drift, in the context of evolving language usage and patterns could pose risks.

Mitigating *model drift i*nvolves continuously updating the model with fresh data to adapt to changes in the underlying distribution. Regular retraining and monitoring of the foundation models are essential to maintain the models' performance over time. In addition to the potentially costly approach of retraining the models (which is also likely beyond the capacity and capabilities of a statistical organisation, augmenting the model with fine-tuning layers based on new and up to date training data may account for model drift, mitigating the generation of errors through this effect.

Handling *concept drift* in large language models requires the ability to detect when the model's performance is degrading due to changing concepts. It may require adapting the model dynamically or incorporating mechanisms to recognise and adapt to evolving language patterns.

Local fine tuning may provide some mitigation against *data drift*, where organisations using LLMs can fine tune the models for their specific use cases, and ensure that the data used in the fine-tuning itself has good consistency between the data used in test and validation procedures, and the production systems. This requires regular monitoring and updates to test and validation datasets.

Parameter Space and Accuracy

The relationship between the parameter space and accuracy in the context of LLMs has similar considerations to the problem in modelling in general.

The parameter space of a GPT refers to the total number of learnable parameters in the model. In general, a larger parameter space allows the model to have greater learning capacity. This, in principle, allows the model to potentially capture more complex patterns and relationships in the training data, which may lead to higher accuracy, especially in use-cases that require a deep understanding of language. However, it is not a simple linear relationship.

Overfitting vs. Generalisation

While a larger parameter space can contribute to better performance on the training data, there is a risk of overfitting. Overfitting occurs when a model learns the training data too well, capturing noise or specific examples that do not generalise to new, unseen data. Due to the lack of intrinsic validation mechanisms, it is difficult to assess whether this balance is appropriately struck in practice with LLMs in general. This is additionally exacerbated in proprietary model offerings because there is no access to the training data, nor the engineering of said data in these instances.

Data Size and Quality, and Computational Resources

The impact of the parameter space on accuracy is closely tied to the size and quality of the training data. A larger model may require a correspondingly large and diverse dataset to fully leverage its capacity. If the training data is limited or not representative, increasing the parameter space is unlikely to result in improved accuracy, and increases the risks of overfitting.

Training and fine-tuning larger models with a vast parameter space, ingesting appropriately sized training datasets demand significant computational resources. While larger models have the potential for higher accuracy, the associated computational costs and infrastructure requirements may render such an exercise inaccessible to various agencies or even governments in general. Smaller models might be more practical in scenarios where resources are limited, especially where the use cases are well defined and the specific validation mechanisms for the limited use case can be established.

Fine-Tuning and Transfer Learning

GPT models can, and often are fine-tuned on specific downstream datasets for specific use-cases (see Section 1.4) to adapt their knowledge to particular domains. The relationship between parameter space and accuracy during fine-tuning depends on the nature of the use-case, the availability of specific data, and the transferability of pre-trained knowledge from the large scale training exercise resulting in the original model. Because of this, careful testing of the fine-tuned models is critical.

Ultimately, the large (and increasing) parameter space can contribute to higher accuracy in GPTs. However, the general balance against issues of input data size and quality, overfitting risks, computational resources, and the like is unclear. Regularisation techniques from broader machine learning approaches, such as batch or layer normalisation are employed in GPTs to aid in decreasing the risks around overfitting and to enhance generalisation. However, the lack of independent metrics of performance, and a lack of robust validation mechanisms exposes model users to the risks discussed above.

Hallucination

"Hallucination" refers to the phenomenon where GPT models generate information that is entirely fabricated or fictional. It can happen when a model generates details, statistics, or events that do not exist in reality. This is concerning, as it can lead to the dissemination of false information or contribute to the creation of "fake news".

Hallucination may occur due to a variety of factors, including:

- **Overfitting to training data**, where a model "memorises" specific examples, rather than general patterns, thus leading to outputs being generated based on the memories, without alignment to input context or prompts.
- **Unintended associations,** where If the training data in a model may contain unintended associations, the language model may inadvertently generate outputs that reflect these.
- **Lack of context understanding**, because GPTs, especially in the context of LLMs do not have an inherent understanding of the context of input data, and is limited in understanding the context of training data, they may generate outputs that are fabricated due to context mismatch
- **Deliberate creative outputs**, where prompting is specifically driven to encourage hallucinatory outputs in order to create "new" content. For example, in the context of LLMs, this may involve prompting the models to output poetry about obscure topics.

Hallucination, in general, adds further risk to the instances discussed prior in the "Accuracy" Section, by potentially creating completely fabricated outputs. It should be noted though that hallucination is used as a general term, in the context of GPTs that captures some of the inaccuracies (e.g., context issues) already discussed.

Measures for mitigating this risk includes*:*

- Regularly validate model outputs against ground truth data.
- Implement strict data preprocessing to remove unintended associations and misleading information from the training data.
- Fine-tune models on diverse datasets to ensure robust generalisation.
- Incorporate mechanisms for controlling the creativity of the model output, especially in applications where accuracy is paramount.

Bias

Bias in LLMs can manifest in different forms, with training data bias and inaccuracy-related bias being two significant aspects to consider.

Training data bias occurs from biases present in the data on which the LLMs are trained. If the training data contains imbalances, stereotypes, or unfair representations, the model will likely learn and reproduce these biases in its generated outputs. This can lead to issues such as:

- Stereotyping, where LLMs learn and perpetuate stereotypes present in the training data, potentially leading to biased or discriminatory outputs,
- Underrepresentation, where certain groups or perspectives are underrepresented in the training data, the model may show a bias towards more frequently occurring patterns, contributing to underrepresentation in generated content,
- And the amplification of existing biases, where biases present in the training data are amplified during the model's generation process, leading to the reinforcement of existing stereotypes and inequalities, particularly if a large volume of content is generated, then published using LLMs with inherent biases.

It should however be noted that these biases are not unique to LLMs, and are potentially present in any algorithm, model, or machine learning approach, where the biases are present in the training data itself.

Inaccuracy biases arise from the LLMs behaving in ways already discussed in the previous subsections, thus generating outputs that are factually incorrect or do not reflect the objective truth. A predominance of such information can be thought of as a type of bias leading to misinformation and the spread of inaccuracies.

Because content generation via the use of LLMs/GPTs are far more time-efficient compared to human generation of content, a large body of misinformation, outdated information, and unverified outputs can overwhelm and outcompete more reliable information on dissemination platforms.

The risk of these types of biases are further exacerbated by the lack of fact checking and verification mechanisms in LLMs in general. Because the human effort required to fact-check the outputs and attempt to correct these are much higher than generating the output in the first place, this can result in an amplification of these types of biases, especially where bad-faith actors may be involved.

For statistical organisations, this highlights a major risk where outputs that are tonally and textually similar to what they may publish, but do not have the same rigour of accuracy checking, are generated and disseminated in channels similar to where they may publish, poisoning the general ability of statistical organisations to provide robust factual information.

4.3. Privacy and Security Concerns

It is important to understand that LLMs share some of the same security limitations as other software applications, as well as having some more security concerns unique to LLMs.

Incorrect application of security controls could allow a malicious user to interfere with the production of LLM outputs or to 'poison' data used in LLM processing.

The best way to use LLMs safely is to understand and plan for the security risks involved and to keep up to date with new security advice as it evolves. *Machine Learning Security Operations (MLSecOps)* is a framework for applying security throughout the LLM life cycle. Where practical, it is recommended that MLSecOps practices are applied when using LLMs for statistical products.

Security risks specific to LLMs generally arise at three points:

1. LLMs can provide an egress point for secure material to leave an organisation.
2. LLMs can be 'poisoned' by the supply of training data that creates adverse outcomes.
3. LLMs can be 'tricked' by carefully engineered prompts into providing confidential information or creating security vulnerabilities.

We will describe each of these in more detail below.

LLMs as egress point

Many implementations of LLMs have components that are stored outside a user's IT environment (for example, calling an external API). In these implementations the model interface sends queries back to its home servers so that it can be processed and then returns a response. It has been shown (https://cybernews.com/security/chatgpt-samsung-leak-explained-lessons/) that sensitive information submitted as part of prompts this way can lead to inadvertent leaking of this information. Further, such queries may become part of the training material for the model, creating an additional risk of exposure of sensitive information in the future, if there are no agreements in place to prevent this happening.

To mitigate this risk, no LLM with externally connected components should be provided with sensitive information. This means not just confidential data but sensitive documents, and code that contains sensitive information such as passwords or other information that should not be placed in the public domain (for example code containing unrevealed parameters for confidentialising data).

Conversely an LLM with access to sensitive information and an outward facing interface may, either through accident or malicious intent, provide that sensitive information to users when provided with certain prompts.

Poisoning LLMs

LLMs can be 'poisoned' by inserting material into their training data that causes adverse outcomes. These could range from impacting the model's performance so it provides incorrect answers to making the model more vulnerable to prompt attacks. A model that has been damaged by poisoning could lead to poor decision making, to reputational risk or to the potentially substantial cost of retraining the model.

Poisoned training data can be used during the original model training or during fine-tuning processes, but the initial training phase is most vulnerable to poisoning. This is because this step often uses large, open source datasets which are much easier to tamper with than the smaller, more curated datasets used for fine-tuning.

One mitigation for this risk is to maintain careful control over datasets used for training. This can be difficult when using pre-trained models where there is no control over which training datasets are used, and worse with commercial models that often supply very little information about what has been used to train the model.

For these reasons the most useful mitigation for this risk is to implement human oversight and checking of any LLM output before it is released publicly.

Prompt injection

Prompt injection is the name given to the use of malicious prompts to obtain undesirable outputs from LLMs. These can include prompts that bypass controls on the LLM in order to provide incorrect, inappropriate or otherwise potentially concerning responses. It also includes prompts that contain code or point to documents that contain code that the LLM will subsequently operate on datasets the LLM has access to. This might include code enabling the release of sensitive information. Prompts can also serve as an ingress point, allowing malicious code access to an organisation's internal IT environment without having to traverse firewalls or other security measures. The range of possibilities for prompt injection is growing rapidly as people gain a better understanding of the behaviour of LLMs, of their implementations and their vulnerabilities.

Mitigation of these risks involves requiring human approval before an LLM can take actions such as running code, sending emails or connecting to external systems. It can also involve building intermediate applications so that end users cannot query the LLM directly (and so that input validation can be applied to prevent malicious code being used), or ensuring the LLM has only minimum privileges and limited access to information (for example, it can only access summary statistics not unit record data).

Applying Privacy Principles and Requirements

NSOs and other statistical bodies need to comply with legislation and privacy standards within their own jurisdictions. Some overarching privacy principles can be applied however:

1. Use only for purposes collected: data collected for a particular statistical purpose (e.g., dissemination of statistical data) should be prevented from being used for other purposes (prepopulation of survey details).
2. Consent: data used for fine-tuning models (and in other training methods) must consider a consent lifecycle (obtaining consent for use of the data, recording consent and the ability to withdraw consent). The responsibility for consent is shared between the LLM model creators for any data used in the initial creation of the LLM and the statistical organisation when using methods, such as fine tuning, for additional training data. It may not always be clear to users whether the LLM output was based on the original training data within the LLM or the fine-tuned data (by the statistical agency), so the more information that can be given by the statistical agency on the data added to the model (and consent obtained) is useful to prevent any misunderstanding to the LLM user.

3. Privacy rights: much like consent, individuals must be allowed the rights of erasure or correction of training data and the right to object to their data being used.
4. Transparency: both internal and public facing statistical use should have as much transparency as possible, from LLM name, source and version used, to additional training data used, to methods used to anonymise data, protect data and any data lifecycle policies.
5. Fairness: When using LLMs for statistical outputs, care should be given to make sure that outputs are not discriminatory. This includes any indigenous data used in statistical outputs.

When using LLMs, consider specific privacy principles for your organisation and how the use of training data affects those privacy principles.

4.4. Copyright and Legal Issues

Authorship and ownership of content used by or created by LLMs is likely to be an ongoing issue with large language models, as it has also been pointed out by the OECD[16]. Some LLMs have been trained on massive amounts of data that includes copyrighted data, in some cases without authorisation or consent. LLMs can also generate output that is very similar to original content, making it more likely to come under legal challenge. In many cases LLM providers did not provide a full disclosure about the data used in training, making it difficult to assess the impact of copyright violations.

A number of mitigations have begun to appear in response to copyright issues. While there has only been some limited success in watermarking content and AI detection systems, there is some promise in the take-up of the C2PA protocol[17], which uses cryptography to ensure provenance of original content is available via metadata as well as all subsequent edits to that original content.

Some LLMs have begun to cite sources within output, while some others will cite sources if asked in prompts. Several companies have also created pledges to customers to assume responsibility for any lawsuits arising out of copyright claims, as long as those customers follow guidelines for the use of those LLMs.

It is important for statistical agencies to understand the risks of copyright infringement and to come up with mitigation strategies. As seen in the previous sections, the risk (impact and likelihood) of copyright infringement and other legal issues varies depending on the specific use case. While use cases that statistical organisations would be interested in, such as code translation and analytical uses, might pose lower direct risk when compared to other cases (e.g., writing in a certain author’s style), they might be exposed to indirect risk. For example, if a LLM provider needed to remove copyright material from their model based on a successful legal challenge, this has a high likelihood of changing the behaviour of the model and may mean that the statistical agency will need to move to a different LLM or reverify any outputs produced by the LLM.

[16] Initial policy considerations for Generative Artificial Intelligence
[17] https://c2pa.org/specifications/specifications/1.3/index.html

On the other side, the current policies in most jurisdictions do not allow the registration of works produced by AI for copyright purposes.

4.5. Lack of Literacy and Understanding - Overuse and Misuse of LLMs

As people and organisations increasingly integrate LLMs into their daily operations, one of the points of concern is the lack of literacy and understanding about these powerful tools among employees and managers, as well as the absence of official guidelines on the matter. There are several risks associated with the uninformed and uncontrolled use of LLMs for daily operations that can be classified in two main categories: confidential information inadvertently being leaked, and erroneous information accidentally being introduced in the working processes.

In the absence of official guidelines, employees may leverage LLMs to perform their tasks, often without a clear understanding of the potential downstream impact of their actions. This has the potential to parallel the shadow IT phenomenon of the past, where employees used non-corporate IT tools without oversight. In this case, LLMs used outside the perimeter of official guidelines may become the "shadow AI", leading to potential security and reputational vulnerabilities.

Employees may inadvertently expose sensitive or confidential information when using public AI cloud services. The perceived confidentiality of a chat environment may mislead the user about the privacy of the input provided, and there are already a number of cases of confidential information leakage in large organisations. This risk is compounded by the fact that user conversations on public AI services are often used to train the future LLMs generations, effectively consolidating the information leakage.

Moreover, LLMs tend to generate text which inspires confidence in the reader - as described in Section 3.1 - even when producing erroneous output. Reliance on LLM-generated output, coupled with a lack of critical evaluation, represents another significant vulnerability. Employees might place unwavering trust in LLMs, failing to validate the accuracy of the generated content. This can introduce errors or biases in the working processes, and ultimately lead to publishing incorrect information in official documents.

Undeclared use of content generated by LLMs may also expose the organisation to reputational risk if the content is included verbatim into official documents without being explicitly declared as such.

To address these risks effectively, organisations must prioritise the education of their employees and managers regarding LLMs. This includes technical and non-technical training, encompassing the technology's capabilities, limitations, ethical considerations, and validation processes. By fostering LLM literacy among employees, organisations may empower their teams to harness the potential of these tools while minimising the associated risks. Clear guidelines and policies should also be established, striking a balance between allowing for responsible LLM usage and maintaining organisational control, recognising that a total ban on LLMs use may not be realistic nor desirable.

CONSIDERATIONS

05

5. Considerations as Statistical Organisations Move forward with LLMs

LLM offers many opportunities for statistical organisations, but it is crucial to proceed with caution while taking various factors into account when integrating LLMs within the organisations. 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 5.1, 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 LLMs

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 organisational 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 LLM/GPT 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 organisations will depend on how the LLM/GPT is entering into the organisation'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 organisation 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 organisations 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 organisations 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 thetop 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[18].

Given these conflicts and tensions, we do not recommend banning LLM/GPTs as this creates a risk of shadow AI with statistical organisations. 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 organisational 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 plagiarising existing publications - while computer-generated text are still to be finalised, 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.

5.2. Engagement with Tech Companies Who Provide LLM 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 organisations to also explore and emphasise 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 organisations can effectively navigate this space.

[18] https://crfm.stanford.edu/fmti/

Role of cloud providers

Cloud service providers are integral to the operation and advancement of LLMs. When engaging with these providers, statistical organisations 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 organisation are also crucial considerations.

The global cloud market is primarily dominated by the 'big three': Azure, AWS and Google Cloud. However, alternative providers often specialise in niche services that offer specific integrations, potentially more suitable for certain statistical organisations. 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.

LLM 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 organisations, recognising and understanding these multifaceted roles is crucial. By identifying the specific category or categories a company operates in, statistical organisations 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 utilising efficient training platforms. Furthermore, understanding these categories helps statistical organisations 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 specialising 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 organisations 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 organisations is crucial. This involves compatibility with existing infrastructure and the ability to adapt to specific technical requirements. Alignment with the ethical standards of statistical organisations 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 organisations, 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 utilise 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 Customisation and Fine-Tuning Services is a crucial segment of companies that includes AI startups and specialised technology firms that tailor existing LLMs to meet specific customer needs. Their adaptability and ability to integrate customised solutions into existing systems are key considerations for statistical organisations.

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 organisations 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 organisations 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 organisations 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 organisation 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 utilisation, and the access to inputs and outputs when utilising 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 customise an LLM to particular data and use cases can prove highly relevant for a statistical organisation.

Regarding openly accessible models, diligent examination of the licensing terms is imperative. LLM creators may impose specific conditions governing the utilisation 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 organisations 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 organisations 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 organisations 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 organisations' core business (i.e., production of official statistics and data services) heavily relies on the public trust requires that statistical organisations 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 organisations are interacting with LLMs. After all, the public has a fundamental right to understand how their data might be utilised and to be assured that measures are in place to protect their data.

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

- using LLMs purposefully where there are clear benefits: it is essential to clearly communicate why LLMs are used in statistical organisations 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 organisations 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 emphasising 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 organisational 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 organisation 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 organisations 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 organisations 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 organisation - 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 organisations 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 organisations 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 organisations 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.

Recognising 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 organisations. 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.