Discussion Paper

How to cope with innovation and the pace of change: the case of Large Language Models at Statistics Netherlands

Marc Ponsen
Marco Puts
Vera Toepoel

November 2025

Abstract

Large Language Models (LLMs), such as GPT-5 by openAI, along with earlier pretrained architectures like BERT by Google, have revolutionized the field of Natural Language Processing (NLP). These models, capable of generating, translating, summarizing, and answering questions based on text prompts, have garnered immense interest across various sectors, including government and industry. Like many other organizations, Statistics Netherlands (SN) explored the potential of LLMs. The goal of this report is to report how we dealt with innovation and the pace of change with respect to LLMs. Research was conducted to assess whether LLMs are merely a hype or genuinely useful for responsible usage at Statistics Netherlands. This report outlines these findings.

The research had four objectives. The first objective aimed to demystify the algorithms behind LLMs. The second objective sought to classify different usage types of LLMs within SN, also identifying the complexities and risks associated with each type of usage. A usage type can be defined as a cluster of usages that share similar properties. The deliverables were training materials and (high level) guidelines for responsible usage of LLMs. These deliverables served a single goal, namely to raise *Artificial Intelligence (AI) Acceptance* within an organisation that is comfortable with traditional statistical methods but new to AI/LLMs. The third objective addressed the feasibility of implementing LLMs on SN's internal IT infrastructure. The study focused exclusively on open-source LLMs that can be securely used internally, adhering to government recommendations. Pilot demonstrations were implemented for usage-types that were expected to have a broad impact at SN. The final objective focused on cataloging existing LLM projects at SN and proposing new projects to enhance operational efficiency. The deliverables were hands-on demonstrations, with the goal of engaging in practical, hands-on experience.

We concluded that the application of LLMs may hold significant promise for enhancing the efficiency and effectiveness of SN's statistical processes and operations. By understanding the underlying technologies, identifying practical applications, and adhering to responsible AI guidelines, SN can safely experiment with LLMs to improve its services while mitigating associated risks.

We advise that future efforts should focus on knowledge dissemination, targeted training (or acquisition) of personal, strategic positioning of SN in the landscape of generative AI, and above all gain practical experience with LLMs. In terms of concrete projects with LLMs, it was advised to roll out a chat- and code-assistant for all SNs employees, in particular for standard business operations. Additionally, it was advised to (continue) research into more niche applications with LLMs.

Contents

1 Introduction

Recent developments in AI, specifically in the field of Large Language Models (LLMs), present opportunities for Statistic Netherlands (SN). There are many applications for official statistics. SN lacks detailed knowledge about this new technology to determine if it is suitable for application in official statistics. Therefore, it is important to i) acquire knowledge about these developments and ii) investigate the possibilities for responsible use within official statistics. Therefore, in September 2023, an investigation was initiated into the potential use of Large Language Models (LLMs) at SN.

On 11 December 2023, the Dutch government published its preliminary position on the use of generative AI (gen-AI) by governmental organizations. Gen-AI is seen as a 'promising and advantageous development to work more efficiently or achieve public goals more effectively.' At the same time, it is acknowledged that there are risks associated with the application of gen-AI within the government. The policy letter further elaborates on the difference between non-contracted and contracted gen-AI. Non-contracted services include publicly accessible forms of gen-AI developed and offered by third parties, such as ChatGPT. The stance is that non-contracted services generally do not demonstrably comply with applicable privacy and copyright legislation, and their use by national organizations is therefore, in principle, not permitted. Concerns lie with issues of leaking private information and unintended training of commercial models. Open-source LLMs are pretrained models that can be downloaded from the internet and used on own IT infrastructure. These may fall outside this categorization, as recognized by the Dutch government, and are not excluded to experiment with.

Commercial and open-source LLMs have both advantages and disadvantages. The former are managed by large technology companies (such as OpenAI's ChatGPT) and can be used for a fee. The advantages of commercial LLMs are that they are extremely powerful models that work robustly and quickly on expensive IT infrastructure. The disadvantage of commercial LLMs is that there are data security concerns and insufficient insight into how the models were trained. Open-source LLMs, on the other hand, can be used in the organization's own IT infrastructure, and data security is in accordance with the desired standards of the organization. Disadvantages lie in the fact that they require investments to get them running internally and they might not be as effective as commercial models. It should be noted that (national and international) policy on AI and LLMs was very much fluid at that time, but it was decided that the scope of SN research was predetermined by the government letter to focus exclusively on open-source LLMs that could be utilized internally. Commercial services, such as ChatGPT or Co-Pilot, were thus outside the (main) scope of SN research.

A critical note has to be made on open-source LLMs: what constitutes an LLM to be open source is still very much open to debate. Many factors may be considered, e.g., is the source code shared, is the algorithm structure and training procedure known, are the model weights shared, and are the training data openly available? The discussion about the desired degree of openness of LLMs is still ongoing and extends far beyond the walls of SN. For the current research, it was decided that a pre-trained model that could be used internally was appropriate given the regulations.

The research had four objectives. Firstly, a prerequisite for the successful implementation of LLMs (or AI solutions, in general) within an organization like SN is increasing knowledge about these new methods. LLMs are the result of decades of research in Natural Language Processing and Machine Learning, and fully comprehending these algorithms is far from trivial. By raising 'AI

awareness' and 'AI literacy' within the organization, better assessments can be made about where opportunities lie and how the techniques can be responsibly deployed. In particular, detailing the differences, advantages and disadvantages of AI models as compared to traditional statistical methods is an important step towards the acceptance of AI in the statistical process of a National Statistical Institute (NSI).

Secondly, it was observed that LLMs can be used in various ways, each with different complexities and risks. A common task, such as programming using LLMs to create or document code, is less risky compared to having LLMs doing automated calculations in a production pipeline. There is a need for SN to develop a vision of what constitutes 'responsible use of LLMs' (often referred to as 'Responsible AI') given various usages.

Once it is decided that a certain type of LLM use can be responsibly implemented in the statistical process and/or daily operations, the third objective is whether CBS can technically facilitate this. As mentioned above, the SN research focused exclusively on open-source LLMs.

Finally, the intention is to actually implement LLMs into people's daily workflows. Projects involving LLMs should enable SN employees to gain experience with various applications. Ultimately, successful projects will be the ultimate proof that LLMs secured a place in official statistics.

To conclude, the goal of this investigation is to report how we dealt with innovation and the pace of change with respect to LLMs, to assess whether LLMs are merely a hype or genuinely useful for responsible usage at Statistics Netherlands. In the next paragraphs we will discuss in more detail the findings of the research. In Section 2 we first discuss initiatives at SN to raise the so-called **'AI-awareness'**, in this particular case targeted to LLMs (covering the first and second objective). Next, in Section 3 we discuss several pilot **demonstrations** of LLMs on SN infrastructure (covering the third and fourth objective). Finally in Section 4 we conclude and suggest future work. Among others we report on ongoing LLM projects at SN.

2 AI Acceptance

Recent advancements in Artificial Intelligence (AI), such as Large Language Models (LLMs) are revolutionizing the world and organizations in many different sectors (e.g., in education, healthcare, government, commercial and more) are searching for AI applications in their work processes. The same holds for NSIs. However, for NSIs the adoption of AI is not straightforward for several reasons. Arguably the most important reason is that governments and policy-makers rely on official statistics to be accurate . Yet many of AI based systems produce outputs that may may be flawed (e.g., due to biases in data or misuse of algorithms) and/or poorly understood (in particular for black-box algorithms, these typically lack the quality standards of traditional statistical methods). The problems of validity and explainability prevent AI methods from being broadly accepted. The successful integration of AI in official statistics hinges on addressing two crucial steps: increasing AI literacy and ensuring responsible usage of AI at Statistics Netherlands.

2.1 Increasing AI Literacy

Increasing AI literacy is essential for the widespread acceptance and effective use of AI technologies in official statistics. AI literacy encompasses understanding basic AI concepts (and

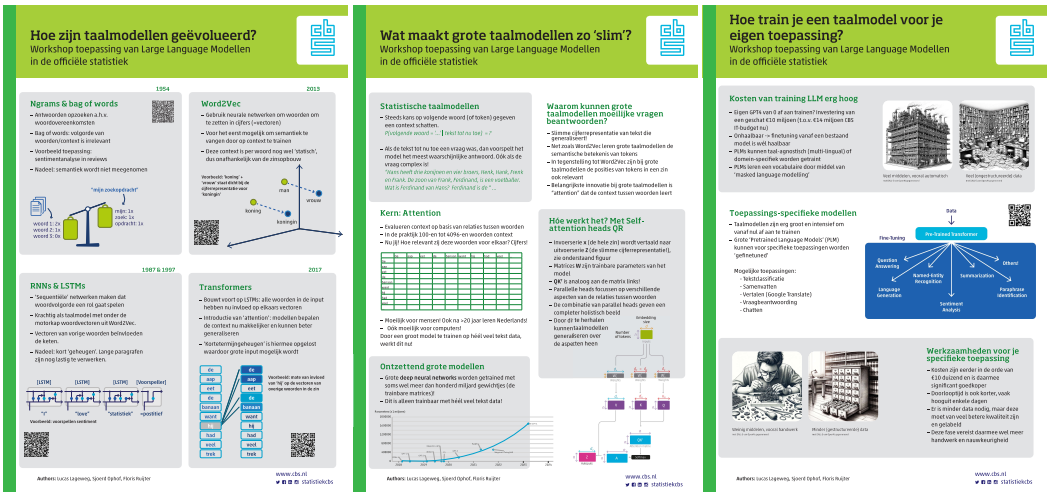

Figure 2.1 Three posters for basic understanding of Large Language Models focusing on evolution in Natural Language Processing, the workings of LLMs and their potential applications in Official Statistics.

comparing them to concepts in traditional statistical methods), recognizing AI's capabilities and limitations and developing skills to work with AI tools. During SNs study into LLMs, a literature study gave the organization a basic understanding of the underlying mechanics and training process of LLMs. This knowledge was disseminated in the organization via workshops, training programs, lectures and via poster(presentations) (see Image 2.1). Knowledge on LLMs or AI in general is crucial for the next step towards AI acceptance, namely using this knowledge to setup guidelines for responsible usage of AI (e.g., LLMs).

2.2 Ensuring Responsible Usage of AI

Responsible AI (RAI) is a widely recognized concept. RAI guidelines are published online by many organizations (e.g., bigh-tech companies such as Google, Meta, Microsoft and more) and it is an actively studied topic in academics. RAI refers to the development and deployment of AI systems that adhere to ethical principles and guidelines to ensure fairness, transparency, accountability, and respect for user privacy. In the Netherlands RAI has received significant attention and this paid off. In an extensive assessment of country commitments to RAI, the Netherlands ranked first [1)]. In short, typically these are the mentioned pillars of RAI:

- **Fairness**: Ensuring AI systems are free from biases and provide equitable outcomes for all users.
- **Transparency:** Making AI processes and decisions understandable and explainable to users and stakeholders.
- **Accountability:** Establishing mechanisms for oversight and accountability, ensuring entities can be held accountable for AI systems and their impacts.
- **Privacy:** Protecting user data through stringent data privacy measures and secure handling of information.
- **Robustness**: Ensuring AI systems are reliable, secure, and perform consistently across various conditions and scenarios.

1) According to the 'Global-Index on Responsible AI, see https://global-index.ai/Countries

When a black-box AI model (such as an LLM) is used in production, all pillars should be considered and assessed. Part of the research therefore focused on setting up practical guidelines for the safe usage of LLMs. The 'Privacy' aspect was dealt with a priori, by focusing solely on open-source models that can be used on on-premise computers (fully disconnected from the outside world, with all of SNs safety protocols in place).

One observation is that LLMs can be used in many different ways and that the risks per type of usage may differ extensively. When using an LLM for code-generation, one can imagine that more lenience is allowed compared to having LLMs having a direct impact in the production process of statistics.

The goals were twofold: (1) identifying types of LLM usage within SN, and (2) linking guidelines for responsible usage for each of the identified types. A usage type can be defined as a cluster of usages that share similar properties (e.g., common tasks such as meeting transcription and summarizing, code-generation and tasks specific to the statistical process such as classification of text, imputation and dissemination). Based on desktop research of internal and external sources, classifications of LLM usage types were created. These classifications may inspire employees at SN on how to use LLMs in their daily work. The idea was to setup practical guidelines for RAI for each individual type. [2)]

Ideally, every project at SN intending to utilize LLMs would be classified into a specific usage type, and as a result the associated risks and conditions for responsible usage would be directly known. In practice, however, an all-encompassing classification proved too complex (i.e., there was always some nuance that did not fit well). Eventually, it was decided to write up general guidelines for RAI tailored to LLMs. This approach aligns with what other organizations have done by establishing principles for the responsible use of AI.

In conclusion, the integration of AI into official statistics presents both opportunities and challenges. By focusing on increasing AI literacy and ensuring responsible usage, statistical agencies can harness the power of AI to enhance their operations while maintaining ethical standards and public trust. These steps are vital for fostering an environment where AI technologies can be used effectively and responsibly to advance the field of official statistics. The study into LLMs has yielded a basic understanding of LLMs, materials (e.g., training program, poster presentations) for dissemination of that knowledge and finally rudimentary guidelines for safely using LLMs in the statistical process.

As a final remark, any adoption of AI within official statistics must remain fully aligned with the European Statistics Code of Practice (ESCoP, Eurostat 2018). The ESCoP principles provide a normative framework within which AI must operate to be responsible in the context of official statistics. These principles apply equally to AI-based systems as well as to traditional statistical processes, as shown by Dumpert et al. 2025 and by Puts and Daas 2021. Embedding the ESCoP explicitly into the assessment and governance of AI ensures that the application of AI guards the trustworthiness of official statistics.

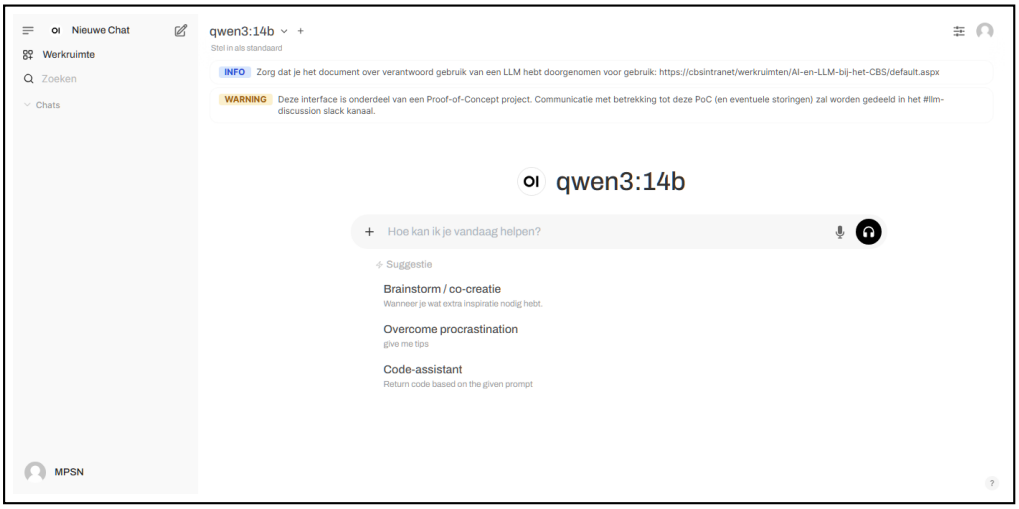

Figure 3.1 The internal chat-assistant allows SN employees to perform basic operations with LLMs such as summarizing of texts, co-creation of texts and searching for information.

3 Projects with LLMs

In the previous section we outlined work with the intended goal of raising so-called 'AI Acceptance' within SN by providing support, training and rudimentary guidelines for safe usage of, in our case, LLMs in the statistical process. Ultimately, successful projects will be the proof that LLMs secured a place in official statistics. In this paragraph, we will first describe the choices made for IT-infrastructure and choice for LLM models. Next, we will highlight the various projects with LLMs at SN.

With respect to hardware, projects with LLMs were done on SNs GPU cluster. The software to run LLMs is extremely complex and organizations such as SN are forced to start with already existing open-source models and frameworks. Reinventing the wheel should be avoided. Most popular programming languages offer a wide variety of open-source software. Therefore, desktop research was done into open-source LLM-models and -frameworks for the programming language of Python. The most popular models and frameworks that aligned with the requirements were selected for experimentation, such as a front-end UI for LLMs (for offering ChatGPT-like chat functionality) and various LLM-frameworks (or packages).

LLM models were selected based on various benchmark scores (i.e., tests to indicate the quality of pre-trained LLM models across a wide variety of tasks. This typically involves running tests on large datasets that contain ground truth values). The models were selected that offered the best performance, while still being feasible to execute the model on SNs GPU cluster.

Typically, a larger (in terms of trainable parameters) LLM performs best, but requires more computer memory and processing power to run efficiently. The final selection criteria was that the model was open-source. Several relevant observations could be made with the model selection. Firstly, the top of the ranking was dominated by commercial models. Secondly, the best ranked open-source model at the time of the research (December 2023) was ranked 7th. At the time of writing this paper (June 2024) it dropped 90 places. This is an indication how quickly

2) See Section 3.2 for some promising usage types at SN.

the world of LLMs is evolving and how difficult it is to be on top of your game. A final observation is that these benchmark scores are difficult to interpret, i.e., it is unclear if commercial LLMs would significantly outperform open-source LLMs in use-cases relevant to SN. With respect to LLM-frameworks, it was decided to interface with the popular website HuggingFace.

We now discuss the various use-cases and projects that were done at SN. The use-cases can be roughly divided into two categories, namely (1) a chat- and code-assistant for general purposes and (2) niche applications of LLMs in the statistical process.

3.1 Chat- and code-assistant

One of the most important conclusions of the preliminary research into LLMs and debates with experts and management was to facilitate the safe usage of LLMs via a chat- and code-assistants. This conclusion was made because the biggest impact of LLMs (initially) is expected to be in the area of AI automatization, i.e., having an AI doing time-consuming and repetitive tasks. Among these repetitive tasks most certainly fall the drafting (e.g., of project proposals), transcription of meetings and the summarizing of large bodies of text. Automating (or assisting in) such tasks can save SN huge amounts of time. In terms of RAI, it is key that people take ownership of the output generated by the model.

The chat-assistant is versatile and can be used for nearly all identified use-cases such as answering questions, generating code, or summarizing texts. The chat-assistant requires manual input of the user, therefore it can not be used automatically into production processes. In other words, human oversight is guaranteed. This makes it an ideal application for wide deployment at SN, allowing the organization to safely experiment with LLMs. See Figure 3.1 of the current solution that is used widely at SN.

The code-assistant addresses a niche task, namely aiding programmers in generating, translating and documenting code. Commercial tools such as GitHub Co-Pilot are widely used by programmers world-wide. A large part of coding at SN is done in secure (internet-disconnected) environments. The main difference with the Chat-assistant is that the Code-assistants are integrated in various programming tools (e.g., VSCode, R-Studio) via a plugin.

It should be noted that these implementations (for both assistants), during the initial research in the beginning of 2024, were not fit for roll-out over the entire organization (due to hardware and software-management issues). For that reason an IT-project started mid 2024 where the focus was placed on two specific use cases: the chat assistant, aimed at all SN employees, and the code assistant, targeted at software developers. Based on the functional requirements, various LLMs were tested, with the size and complexity of the models significantly influencing the quality of the response and system performance (response time). During the Proof of Concept, test groups were established: 50 people for the chat assistant and 20 for the code assistant. In addition, two special load test days were organized to test the capacity of SNs GPU cluster under load. These tests provided insight into the system's performance under high load conditions. The tests provided both quantitative (e.g., on response latency of the assistants) and qualitative results (e.g., users were send a questionnaire).

The test results were generally successful or partially successful (see Figures 3.2 and 3.3). Users were generally satisfied with the availability, reliability, and overall quality of the assistants. Given users' familiarity with high-quality offerings from major tech companies, there was a key concern they might find internal alternatives lacking — but this did not (really) appear to be the case for the tested use-cases. Therefore, the overall conclusion of the IT-project was to proceed

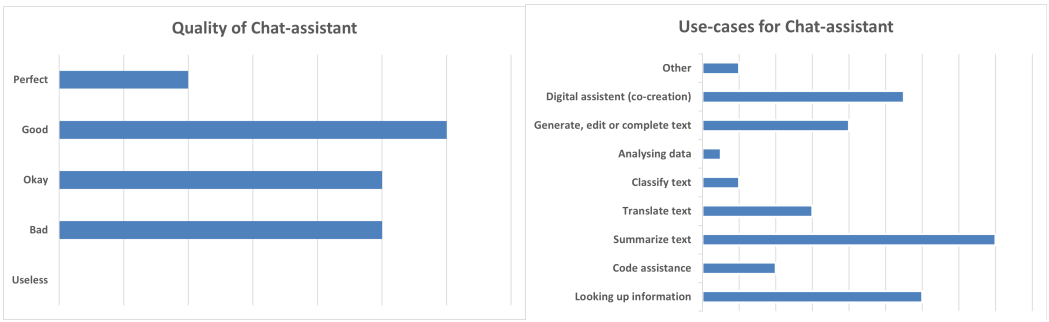

Figure 3.2 Results for the proof of concept with the chat-assistant.

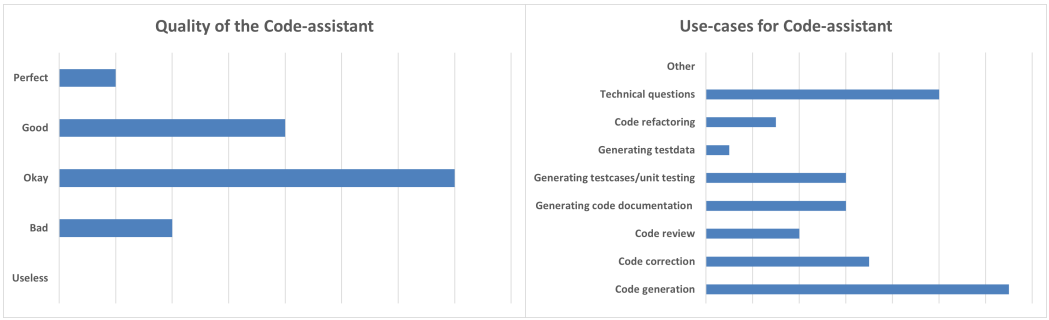

Figure 3.3 Results for the proof of concept with the code-assistant.

with the implementation of these assistants. The advice involves the acquisition of scalable new hardware to host the assistants and delivering a platform service (maintained by the IT department). Additionally, Functional Managers from the business side (typically experts in the field of LLMs) should be made responsible for configuring and managing the LLMs.

3.2 Niche applications and research with LLMs

The previous paragraph provided more insights into SNs approach for facilitating (generic) LLM-assistants for all SN employees. In the current paragraph we will describe several niche LLM initiatives that followed the initial research results. These initiatives can be roughly divided into several categories, namely (1) fundamental research into Responsible AI, (2) LLM improved search functionality, (3) classification of free text into statistical concepts with LLMs and (4) LLMs in the information dialogue. We will provide more detail for these four categories below.

1. **Fundamental research into Responsible AI** As part of SNs multi-year methodological research program, a new research track was launched called ”Applicable AI”. This research theme supports efforts to responsibly integrate AI into statistical processes at SN. The successful use of AI in official statistics presents challenges that will be addressed from three different perspectives: normative, methodological, and application oriented. The normative framework focuses on the overlap between the *European Statistics Code of Practice* and *Responsible AI* principles (as mentioned before) and conceptualizes standards for AI when using it in official statistics. Key issues include the transparency and quality of AI models. The methodological framework addresses how to meet these standards, covering representativity, validity, explainability, and bias. Finally, practical implementation poses its own challenges, as AI applications typically need to be tailored to specific problems. Furthermore, a PhD study is underway to assess the suitability of pre-trained LLMs for use in different statistical domains, with a focus on quality indicators that measure how well LLMs grasp statistical concepts.

2. **LLM improved search functionality** A key area of ongoing LLM research at SN is aimed at improving how easily our data can be found and accessed. As for example, an AI-powered data search engine was created for Statline tables, that matches user queries with semantically enhanced meta-data using LLMs. Similar studies were also done for SNs internal data (in contrast to Statline, which only involves SN's external publications), more specifically improving the search for micro data and internal documents or reports.

3. **Classification of free text into statistical concepts with LLMs** A LLM use-case that is expected to have huge impact is having LLMs classify unstructured, large bodies of text. Some examples that were investigated during the research include classifying skills in vacancy texts, classifying the "Standard Industrial Classifications" code based on descriptions in the business registry, classifying various codes of different accounting tools into the standard "Reference Classification System of Financial Information", classifying feedback on questionnaires into a fixed number of categories and more. All these tasks involve taking as input a large body of text and then relate this to a fixed set of (statistical) concepts or categories. This use-case typically requires automatic access to LLMs (e.g., via programming code, rather than via a chat interface). Taking into consideration that this task may have a significant and direct impact to the production of statistics, only a small group of experienced data-scientists experimented with this use-case (taking the guidelines for responsible AI into account). Research into these applications is still ongoing, nothing has made it to production yet.

4. **LLMs in the information dialogue** A relevant task for a National Statistical Institute is the dissemination of statistics and information. Dissemination of statistics is typically via tables, charts and supporting articles. In recent years, SN also investigated a more interactive way of dissemination statistics via question-answering systems. This 'information dialogue' can also be addressed having LLMs answering questions based on a large number of text documents. Technically, this is known as Retrieval-Augmented Generation (RAG), which involves extracting facts from an external knowledge source. Several (ongoing) projects at SN (together with academia) look into LLMs to find the desired information in Statline tables and disseminate it via text (see Jonas Kouwenhoven 2015) or charts.

4 Conclusion

Recent developments in AI, specifically in the field of Large Language Models (LLMs), present opportunities for Statistic Netherlands (SN). There are potentially many applications for official statistics with far-reaching implications. SN lacks detailed knowledge about this new technology to determine if it is suitable for application in official statistics. Therefore, it is important to i) acquire knowledge about these developments and ii) investigate the possibilities for responsible use within official statistics. Therefore in September 2023, an investigation was initiated into the potential use of Large Language Models (LLMs) at SN. The objectives were to raise so-called 'AI Acceptance' at SN, firstly be increasing knowledge on LLMs and secondly by providing user practical guidelines to use them safely in work.

The second part of the research targeted hands-on experience with open-source LLMs. Several types of usages were implemented on SNs GPU cluster, as for example a general purpose chat- and code assistant. This is an ideal application for allowing the organization to safely experiment with LLMs. A proof-of-concept with both assistants yielded in positive feedback. Also, several niche applications of LLMs was experimented with. For these use-cases, some of which are more

closely integrated in the production of statistics, it was advised to study this with selected, well-trained people and ensuring that RAI guidelines are followed.

The following two suggestions for future are as follows.

Firstly, it is crucial that SN clearly defines and firmly establishes its position within the landscape of large language models (LLMs). SN is an active participant in a major international consortium of national statistical institutes (AIML4OS), funded by Eurostat, and collaborates with other national governmental agencies on AI and LLM-related initiatives. At the time of writing this paper SN sees itself as a 'smart follower'. Basic understanding and usage of pre-trained, open-source LLMs is considered, but SN does not have the ambition to eventually train (statistical) LLMs (or 'foundation models' in general). SN may implement niche applications with LLMs, but this includes conducting fundamental research to validate or explain the outcomes of an LLM, successfully training and/or constraining the output of an LLM, or gaining expertise in state-of-the-art LLM frameworks and GPU computing clusters.

Secondly, future efforts should focus on knowledge dissemination, targeted training (or acquisition) of personnel, and, above all, gaining practical experience with LLMs in a responsible way. The suggestion therefore is to communicate the guidelines for RAI internally and have these guidelines under continuous scrutiny. Additionally, SN's current guidelines for RAI are abstract and high-level. In the end, statisticians and data-scientist need hands-on, practical rules and best-practices for safely using LLMs. This in part is an organizational challenge (e.g., in sharing knowledge and user-experience, implementing human oversight and more) but it also holds considerable methodological challenges (e.g., in validating and explaining output of LLMs). With respect to the latter it was decided to focus on RAI explicitly in SNs research program for the next five years.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague, Netherlands
www.cbs.nl

Prepress
Statistics Netherlands Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information