Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

26 April 2024 V1.0[1]

Fact sheet on raising awareness of large language models in the Federal Administration

What are large language models?

Large language models (LLMs) are AI technologies that form the basis of the most common AI systems[2] in the field of natural language. They are the core of many generative AI tools.[3]
LLMs are able to process, understand, interpret and generate natural language and can perform a variety of tasks such as translation, text comprehension and text generation. Trained on large volumes of data in many different formats, the latest LLMs can produce texts that are often not easy to distinguish from texts written by humans. Even the input of sound or images is conceivable, as these can now be converted into text, in many cases almost flawlessly. Acoustic voice output is also practically indistinguishable from human speech. Some LLMs are already expanding into “multimodal” models capable of processing and generating not just text, but also images and videos.[4]
Examples of LLMs are the GPT model series (used in OpenAI's ChatGPT and Microsoft's Copilot), Gemini (used in Google's Gemini, formerly Bard), Meta's LLaMA models, X's Grok model series and Anthropic's Claude models.[5] LLMs acquire the ability to interpret and generate natural language for general purposes by “learning” statistical relationships from text documents during a computationally intensive iterative training process. These statistical models are based on techniques and methods for natural language processing (NLP[6]) which enable them to extract meaning and correlations from human language.

Principles for large language models (LLMs) in the Federal Administration[7]

As a user of generative AI technologies (such as LLMs) in the Federal Administration, you are requested to observe the following principles for responsible experimentation. Generative AI technologies (such as LLMs) can support you in your day-to-day work within the administration. LLMs are able to perform a variety of partially or fully automated text-based tasks. Give them a try and learn something new! With a little creativity you can make your contribution to an innovative administration. Caution and critical thinking are advised however – and you must ensure you comply with the regulations at all times.

1) You know what generative AI is and what LLMs are. And you know their limitations.
2) You use generative AI technologies (such as LLMs) lawfully, ethically and responsibly. You check that there is a sufficient legal basis for the intended use. Requirements for a legal basis vary depending on

[1] This fact sheet is reviewed regularly to take into account new developments and an improved understanding of generative AI technologies (such as large language models) in the Federal Administration.

[2] An “AI system” is a machine-based system that, for explicit or implicit objectives, infers from the input it receives how it can generate output, such as predictions, content, recommendations or decisions, which can influence physical or virtual environments. AI systems can be equipped with varying degrees of autonomy. AI technology refers to individual functions that can be implemented in computers to enable AI capabilities (e.g. “machine learning”). An AI system thus refers to a structured, context-bound combination of AI technologies for the purpose of achieving AI capabilities. *Source:* https://cnai.swiss/en/products/terminology/

[3] “Generative AI” is a broad term that refers to AI systems that are trained on large amounts of data from the physical and virtual world in order to generate data themselves (e.g. texts, imagery, sound recordings, videos, simulations, and codes). They are often multimodal, e.g. with input and/or output in one or several modalities (e.g. text, image or video). *Source:* https://cnai.swiss/en/products/terminology/

[4] These are also known as large multimodal models (LMMs) with generative capabilities (https://huyenchip.com/2023/10/10/multimodal.html).

[5] See “Fact sheet on the use of generative AI tools in the Federal Administration”; https://cnai.swiss/en/products-other-services-instruction-sheets/

[6] “Natural language processing” (NLP) is a branch of AI that deals with the analysis, understanding and generation of written and spoken words and sentences (natural language). Most NLP techniques and methods are based on machine learning, extracting meaning and context from human language. Areas of application include text recognition, speech recognition, bots, chatbots and digital assistants. *Source:* https://cnai.swiss/en/products/terminology/

[7] Inspired by the “Generative AI Framework for the UK Government” (dated 18 January 2024); see https://www.gov.uk/government/publications/generative-ai-framework-for-hmg

whether you are using generative AI technology (such as LLMs) to help with a translation, to provide greater support when writing texts or for a fully automated decision.

3) You know how to keep the use of generative AI technologies (such as LLMs) secure.[8]
4) You ensure meaningful human control at the right stage (from development through to the application of generative IA technologies such as LLMs).
5) You are transparent about using AI in general and LLMs in particular. Text created by LLMs may appear to have been written by a human. As a federal employee, you have a special responsibility in both internal and external communication. Labelling of AI-generated text is recommended (e.g. “This text has been produced with the help of AI, but edited and checked by a human.”). For automated, individual decisions as defined under Art. 21 DPA[9] labelling is compulsory.
6) You apply these principles together with the guidelines on AI for the Federal Administration and the Federal Administration's core principles for human-centric and trustworthy data science (and AI) and ensure the necessary (quality) assurance.

Limitations of large language models

LLMs create their output/results based on the statistical probabilities of the next word in a sequence and do not consider whether the output is true. This can lead to factual errors and hallucinations, i.e. output/results that do not correspond to reality (“hallucinations”). This can be limited by providing the model with a context (e.g. a set of text documents) for finding the answer. Nevertheless, hallucinations can still occur if the answers contain content outside of this context. Other limitations of LLMs in the context of proper use include, for example:[10]

- unwanted output/results, word-for-word repetition and bias in the model resulting from the unbalanced composition and content of training data (e.g. personal or copyrighted data and texts with dubious, erroneous or discriminatory content);
- outdated results, because LLMs generate text based on the training data they have processed, which are inevitably limited to content that existed when the respective model was trained;
- irreproducible results, because even if an LLM repeatedly receives the same input, the output/result generated can vary in terms of language and content every time;
- insecure generated code (if an LLM has been trained on code, it can also generate code that will potentially contain security vulnerabilities when executed);
- faulty reaction to specific input as even small deviations in the input can lead to large differences in the generated output/results. If the input for LLMs is different to the texts that were used for training, the model is often unable to process these properly and generates faulty output/results;
- LLMs usually generate texts that are free of linguistic errors and whose content is convincing. They are able to generate meaningful output in a wide range of subject areas. This can give the impression of human-like capability, leading to too much trust in the model's output and ability (“automation bias”);
- the entered data are often reused as training data (this will include objectively incorrect input from users);
- LLMs are prone to interpreting text as an instruction, as they interpret all input in the same way, making no distinction between instructions and other texts;

[8] For their use within generative AI tools, please see the “Fact sheet on the use of generative AI tools in the Federal Administration”; https://cnai.swiss/en/products-other-services-instruction-sheets/

[9] https://www.fedlex.admin.ch/eli/oc/2022/491/de#art_21

[10] The list of limitations is based on Chapter “2.3 Risks of LLMs” from a report (version 1.1 dated 27 March 2024) by the German Federal Office for Information Security on the opportunities and risks of LLMs; see https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf

- the data entered lack confidentiality as LLMs are often provided as an internet service using suitable interfaces (e.g. a web browser). Passing on personal data to such services via the internet may constitute a transfer of data to third parties, which is subject to data protection legislation;
- the development and operation of LLMs by providers on their infrastructure may lead to a high level of dependency (e.g. lack of data sovereignty, lack of control over the model);
- LLMs sometimes generate sources that are not always correct or do not even exist, adhering to them with persistence and conviction.

Generative AI technologies (such as LLMs) offer a wide range of opportunities and possible applications and are currently developing at a rapid pace. This raises new security concerns about the development, operation and use of these models. To use these AI technologies securely, the risks must be analysed systematically.

With this in mind, before integrating generative AI technologies (such as LLMs) into their own day-to-day work processes, all users in the Federal Administration should also carry out an individual risk analysis. The above-mentioned principles can provide a framework for responsible experimentation.

To answer your questions:

- on AI in the Federal Administration: AI working group and competence hubs in the Competence Network for AI (CNAI)
- on information security and data protection: the IT security officers and data protection officers of your administrative unit
- For specific data science and AI services: DSCC

Further information:

Guidelines on AI for the Federal Administration

The seven guidelines on AI for the federal administration continue to apply: Put people first, ensure regulatory conditions for the development and application of AI, demand transparency, traceability and explainability, define accountability, guarantee safety, actively shape AI governance, involving the relevant national and international stakeholders in the process.

Federal Administration's code of practice for human-centric and trustworthy data science (and AI)

The Code of practice provides the federal government's administrative units with guidance and practical explanations to raise awareness of the core principles of human-centric and trustworthy data science (and AI) and enables them to implement these principles in everyday work. The core principles are: Data and information protection, information security, data security, data governance, non-discrimination, explainability, traceability, transparency, reproducibility, neutrality, objectivity and ethical handling of data and results.

This fact sheet was drafted by the "AI in the Federal Administration" working group in the CNAI in cooperation with representatives from all departments and the FCh.