29 August 2024

European Medicines Agency

Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities

1. Scope

This document provides guiding principles on the use of general-purpose large language models (LLM) in regulatory science and for regulatory activities and provides high-level recommendations to regulatory authorities to facilitate the safe, responsible and effective use of LLMs.

2. Glossary

This glossary is a living glossary and is not intended to be a legal definition of the terms, though the definitions used are sought from legal documents where feasible.

Artificial Intelligence (AI)	A machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments
Chatbot	A computer program that simulates human conversation with an end user. Not all chatbots are equipped with artificial intelligence (AI), but modern chatbots increasingly use conversational AI techniques, powered by large language models
Foundation models	Models that are trained on broad data that can be adapted to a wide range of downstream tasks, such as GPT[1]
Hallucination	Outputs of large language models that while seemingly plausible, deviate from user input, previously generated context, or factual knowledge[2]
General-purpose AI	An AI system that can be used and adapted to a wide range of applications for which it was not intentionally and specifically designed[3] *General-purpose AI is often used as a synonym of foundation models, although some authors see 'general-purpose' as broader than 'foundation models', as it focuses on generic capabilities rather than characteristics of model development. It is the main term used in the AI act.*
Generative AI	AI that can generate new content such as text, audio, images, code, and videos.
Large language model (LLM)	A category of generative AI, focusing on text generation.
Prompt	Text that describes a task that a generative AI model should perform. The prompt includes any text provided to the LLM, but this may not be obvious to the user, e.g. an organisation may offer a chatbot that allows users to write a query, but there may be an underlying prompt

[1] Adapted from Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[2] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

[3] Texts adopted - Artificial Intelligence Act - Wednesday, 14 June 2023 (europa.eu) - Amendment 169

	that is given to the LLM with the user query. E.g.: [hidden from the user: act as a medicines' regulator, reply in a formal way]

3. Introduction

3.1. What are large language models?

Large language models (LLM) are a category of generative AI or foundation models trained on large amounts of text data making them capable of generating natural language responses to an input or request (prompt).

In short, an LLM is trained in two phases: during the pre-training, a large set of text data is used and the statistical correlations between tokens, (units of text or size of words), are transformed numerically. These statistical correlations are called weights. In the fine-tuning phase, the pre-trained model undergoes a supervised learning process using labelled examples to update the weights and improve the model for specific tasks.

While users often interact with LLMs using chatbots, LLMs and chatbots are not the same. Chatbots can be powered by LLMs or some other technology, and LLMs can be used in other programs that are chatbots. For instance, GPT 4 is the LLM, ChatGPT is the chatbot.

3.2. Use of large language models

LLMs are typically used in general or bespoke chatbots. General chatbots allow users to query anything as the computer program serves as a general-purpose chatbot for handling a wide range of queries. General purpose chatbots include ChatGPT or Bing AI, which use open online LLMs. Bespoke chatbots on the other hand are tailored to answer specific questions in a particular domain.

LLMs are also commonly used for information processing automation or knowledge/data mining, whereby a developer provides a standard "locked prompt" to a LLM alongside with raw data. The user then retrieves processed data, or navigates the results but does change the prompt, only the raw data changes. For instance, LLMs can be used to create an application that summarises meeting notes from a committee using a specific format. The prompt is always the same, what changes is the meeting notes provided.

LLMs can also be used in AI virtual assistants. These are applications that react to text, voice prompts or inputs and complete some tasks for the user, such as scheduling appointments.

These tasks are varied and include:

- *Writing assistance*: drafting emails, creating content, proof-reading and rephrasing text.
- *Searching for and summarising information:* sifting through large volumes of information quickly to find relevant data and summarising long articles.
- *Language translation*: translating text in different languages.
- *Education and tutoring*: providing educational support and structuring complex subjects.
- *Proving coding language support*: suggesting code on-demand based on user prompts.

LLMs can be classified according to their availability and accessibility into two main categories: open-source models (freely accessible to the public) and closed-source models (developed as commercial products and often necessitating licenses or subscriptions for use).

Access to LLMs is generally enabled through third parties, which may include organisations providing specialised tools developed upon or integrated with LLM technology. We can categorize LLMs in four ways on the basis of how users can interact with them:

1) third-party, open or closed source, externally hosted, available online: users interact with the language model through an online interface. The interaction is dependent on the interface provided by the third-party and may limit customisation or integration with internal systems.
2) third-party, externally hosted, part of enterprise solutions: the interface can be more comprehensive, designed to be integrated into larger systems. Here, users engage with the LLM within a broader array of tools, potentially encompassing data management, analytics, and additional functionalities.
3) third-party, open source, internally hosted: apart from being hosted internally, it is like the previous category. Users exert greater control over their interaction with the LLM, allowing for customisation of the hosting environment, integration with other systems, and potential modification of the model (if open sourced).
4) (re)trained internally: this category offers a unique level of interaction, entirely determined within the organisation. This allows for extensive customisation, including bespoke interfaces, integration with internal data sources for retrieval augmented generation, and fine-tuning performance. Here, additional data from the organisation can be integrated into third-party LLMs, or LLMs created by external entities can be refined using the organisation's data. Moreover, there is the option to develop and train LLMs within the organisation, leveraging its own data and teams, albeit requiring significant investment.

The ways users can interact with LLMs could affect flexibility, control, resource requirements, and integration possibilities during interaction. It is therefore essential to consider these factors when selecting a model type for a specific line of work.

3.3. Safe and responsible use of LLMs

As indicated above, LLMs possess powerful capabilities for executing diverse tasks including text generation, translation, coding support, and knowledge retrieval. Many of these capabilities may also be applied to support various tasks and processes within the medicines regulatory system.

At the same time, the use of LLMs is not without challenges and risks. LLMs have shown surprising failure at seemingly trivial tasks, returning irrelevant or inaccurate responses – known as hallucinations. In addition, LLMs may have not been exposed to information that answers a scientific or regulatory question, due to the novelty of many tasks performed by regulators (e.g. those concerning new active substances). Finally, validation of these technologies is challenging.

Furthermore, LLMs process a variety of users' input ranging from short paragraphs to entire documents. The potential for the prompts being stored and further processed presents risks in terms of confidential information, data protection and privacy.

LLMs are trained on large amounts of text sources, which may include public sources. Public web-scraped datasets may contain (personal) data which could be inaccurate or contain misinformation. Therefore, having controls in place to address data protection risks is very challenging. Considering that these LLMs store the data they learn from in the form of billions or trillions of weights, rectifying, deleting or even requesting access to personal data learned by LLMs is practically impossible.

Furthermore, if not properly secured, LLM outputs might reveal sensitive or private information included in datasets used for training, leading to potential or real data breaches. **Finally, the output**

can infringe on other legal rights, such as copyright, or have ethical implications, for example if it influences people's freedom of choice or moral values.

Regulatory agencies should therefore empower staff to be able to leverage the capabilities of LLMs only where they can be used in a safe and responsible manner.

4. General ethical considerations

In 2018, the European Group on Ethics in science and new technologies (EGE) published a statement on ethics in the use of AI, robots, and autonomous systems. This group, an adviser to the European Commission on matters of ethics, proposed a set of principles based on the fundamental values laid outlined in the EU Treaties and in the EU Charter of Fundamental Rights. These principles include human dignity, autonomy, responsibility, justice, equity and responsibility, democracy, rule of law and accountability, security, safety, bodily and mental integrity, data protection and privacy, sustainability[4].

The High-Level Expert Group (HLEG), an independent expert group set up by the European Commission, then published the Framework on Trustworthy AI in 2019. The HLEG promotes an approach founded on fundamental rights, focusing on four ethical principles to consider during AI design, development and use: autonomy, prevention of harm, fairness, and explicability[5].

On the basis of these principles, the HLEG proposed seven requirements, whose fulfilment through technical and non-technical methods is needed for trustworthy AI: accountability, human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing[6].

The above principles and requirements apply to all AI systems. Weidinger et al in 2021[7] published a pivotal manuscript on ethical and social risks of harms that focused on the use of and interaction with LLMs, from both a developer and a user perspective. Their paper noted risks of the following:

- **Discrimination, exclusion and toxicity** – social harms may arise from LLMs producing discriminatory or exclusionary text, including content that perpetuates stereotypes, exclusionary norms and toxic language.
- **Information hazards** – harms may arise from data protection infringements, via misuse, unsafe use and leakage of personal data or other sensitive information as well as from unsafe data storage and management practices.
- **Misinformation harms** – harms may arise from LLMs disseminating false or misleading information (including misleading medical information) or from content that incentivises unethical or illegal actions.
- **Malicious uses** – risks may arise from the intentional use of LLMs to cause harm, including fraud, disinformation at scale, code generation for cyberattacks.
- **Human-computer interaction harms** – risks may arise from unsafe practices in the design of an LLM application, such as a chatbot, and interaction with it, including overreliance on LLM outputs

[4] European Commission, Directorate-General for Research and Innovation, European Group on Ethics in Science and New Technologies, *Statement on artificial intelligence, robotics and 'autonomous' systems – Brussels, 9 March 2018*, Publications Office, 2018, https://data.europa.eu/doi/10.2777/531856

[5] High-Level Expert Group on Artificial Intelligence (HLEG) – Ethics Guidelines for Trustworthy AI.

[6] High-Level Expert Group on Artificial Intelligence (HLEG) – Ethics Guidelines for Trustworthy AI.

[7] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. [https://arxiv.org/pdf/2112.04359.pdf]

(automation bias), transparency and accountability shifts, and stereotype perpetuation, e.g. a “female” name for a cooking chatbot.

- **Automation, access, and environmental risks** – risks mat arise from environmental harms, social inequality, human labour displacement, disparate access to benefits of hardware and digital literacy.

Building on these ethical principles, the protection of personal data also deserves particular attention. While this document focuses on the use of LLMs, the *processing*[8] *of personal data*[9] in LLMs can occur during their development, implementation, and use without necessarily being obvious. Thus, compliance with the applicable data protection rules[10] and other legal requirements must be ensured during all stages of the life cycle of LLMs.

5. User principles

5.1. Take appropriate measures to ensure safe input of data

Interaction with an LLM typically starts with a user introducing a prompt, a task or a query for the LLM to address. While prompts may be in the form of a simple question, e.g., ‘explain pharmacovigilance in 100 words’, prompts often use contextual information or require the LLM to use some data the user has, e.g., ‘summarise this text: …’. It is key to be aware of this subtle difference as for some LLM interactions, the LLM uses only information in the prompt, while for other the LLM uses information generated elsewhere.

This first interaction with a LLM is critical to ensure lawful, responsible and safe use. Users must clearly understand the capabilities and limitations of LLMs so that they can harness their potential effectively, while being cautious about potential risks. [11] This awareness will empower them to make informed decisions when interacting with these tools.

There are two key elements to ensure safe use of prompts in LLM chatbots or of raw data in LLM information extraction automation and knowledge mining tools:

- **Understanding how the LLMs are deployed**: namely if they are deployed locally and fully controlled by the organisation or if they are open online models that may reuse the data from prompts.
- **Careful prompt engineering** (or crafting): this is fundamental to avoid exposing sensitive data to LLMs, to ensure a better output and to reduce risk of bias in the output from biased prompts.

Recommendations:

[8] According to Article 4(2) of GDPR "processing" *means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.*

[9] According to Article 4(1) of GDPR "personal data" *means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.*

[10] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation); national data protection laws may also apply; Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC

[11] Confederation of European Data Protection Organisations (CEDPO) – Generative AI: The Data Protection Implications.

1) Actively educate yourself about the LLM application you want to use and adapt your prompts to the level of control the organisation has on the LLM application.

2) Carefully draft your prompt including any text generated elsewhere (not by you) and check the text to avoid inputting sensitive information including personal data, trade secrets, data that is protected by intellectual property law, data where existing contracts impose restrictions on sharing and other secrets like passwords and tokens.

3) Be careful when copy-pasting to a prompt. Hidden text may be introduced in the documents you are copying that can modify the behaviour of the LLM and be harmful to you and the organisation. If the input is raw data, ensure you trust the source.

Apply critical thinking and cross check the outputs

LLMs excel at creating content and it is tempting to use it as is, but this attracts significant risks. This is because LLMs use large amounts of text to establish statistical correlations between words. These statistical correlations can then be used to generate new content meaning that LLMs may output text that appears coherent but that is irrelevant, false, biased or lifted verbatim from a source document.

Therefore, users should treat output text cautiously and use their judgement to decide whether the text is trustworthy/correct, whether it is reliable (did the model follow instructions, is the output relevant?), whether it is fair (is the output biased?) and lastly whether it is lawful (e.g. not copyrighted/plagiarised?). The greater the risk of potential harm is, the greater the degree of scrutiny must be. [12]

Recommendations:

4) Avoid automation bias and keep a healthy scepticism: review all LLMs outputs for veracity, reliability and fairness. Consider the criticality of the use case – is it straightforward and low risk or knowledge-based and sensitive? Adjust the extent of the review of the output.

5) Prevent the risk of copyright violation or plagiarism by redrafting output that is new information to you.

6) If the output used, e.g. a text, is mostly the result of an LLM, consider disclosing the output as the outcome of an LLM.

7) Ask the LLM to confirm its own answer. Ask for sources to corroborate the output; for summarisation/content extraction activities, ask the LLM to use only text from the files/content provided and to quote exact sentences so that these can be checked in the content. Consider additional quality control activities as needed.

8) In coding, use cases syntactical correctness is checked when the code is run. This however gives no assurance on the accuracy of the logical and mathematical content of the code or safe coding practices. You should therefore always review and test the code produced by the LLM.

5.2. Continuously learn how to use LLMs effectively

A responsible use of LLMs requires familiarity with this technology and knowledge on how to interact with them. The use of AI in medicines regulation is primarily to benefit public health and an effective prompt maximises the benefit of using LLMs by generating the most likely reliable output, at a lower cost, both financially as well as environmentally. Furthermore, LLMs are evolving quickly, and risks identified above may change, continuous learning helps prevent a shifting risk profile of LLMs.

[12] Vrije Universiteit Brussel (VUB) – Responsible use of artificial intelligence for research purposes.

LLMs users should embrace continuous learning to ensure that LLMs can be used safely, effectively, and responsibly in regulatory science and medicines regulation.

Recommendations

9) Continuously educate yourself on how to use LLMs, including how to use their settings, to improve the efficiency of the LLMs interaction and reduce the financial and environmental costs.

10) Reach out to training networks and/or centres of expertise inside the European medicines regulatory network if you would like information on how to have (additional) training.

5.3. Know who to consult when facing concerns and report issues

From an LLM user perspective, there are roughly three categories of problems they may face: 1) inherent issues with the data used to train the LLM, for instance, becoming aware that the datasets used to refine the model included personal data or intellectual property, 2) what information can be shared in the prompt, including which LLM can be used to manipulate sensitive information, 3) the reliability, veracity, fairness and lawfulness of the output.

Knowing who to contact in an organisation is critical to ensure:

- That a security risk analysis, threat analysis or data protection impact assessment is conducted if necessary. This will also enable the organisation to identify and implement additional security and privacy enhancing measures.
- That the conditions that led to incidents of sharing sensitive information where they should not have been shared are investigated and addressed.
- That severely biased or erroneous outputs are swiftly addressed and do not pose a risk to medicine regulatory authorities.

The importance of consistently addressing concerns is heightened by the fact that as LLMs undergo continuous updates and refinements to enhance their performance or prevent malicious use, new unintended consequences may arise. User feedback plays a crucial role in this process. Ongoing monitoring, including monitoring for personal data breaches, or severely biased or erroneous outputs, is essential to ensure compliance with general ethical principles described in Section 4. [13]

Severely biased or erroneous outputs include generated text that can have a negative impact in decision-making in regulatory science, financial or human resources, or significantly affect some fundamental right, and/or that may affect the reputation of the organisation. For instance, a summary or translation of a meeting about a contract dispute that changes key text, e.g., from criticism to support.

Recommendations

11) Know who to consult when it comes to security and data protection and reach out to the information security team and/or Data Protection Officer (DPO) if you have data protection concerns.

12) Report incidents or severely biased or erroneous outputs to the appropriate function or team.

[13] See also Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, Xiuzhen Cheng – On Protecting the Data Privacy of Large Language Models (LLMs): A Survey.

6. Organisational principles

6.1. Define governance that helps users have a safe and responsible use

Medicines regulatory authorities have much to gain from the use of LLMs. They are also directly affected by the use that staff make of these models and thus ultimately should make efforts to ensure safe and responsible use of LLMs by their staff.

Establishing governance on the use of LLMS can help users navigate the risks of these models, firstly by providing awareness and guardrails on the use of the different types of LLMs, the acceptable LLM risk profile of the organisation, the LLMs available to them and potential risks encountered from their use (see section 5.1).

Governance may also consider whether mandatory training on safe and responsible use of LLMs is required either for the whole organisation or based on the risk profile of the use cases and how to collect information on errors and issues arising from LLMs.

Another crucial aspect of governance relates to the development of these models, not just their use. While some agencies might not use their own data to fine-tune or improve LLM performance on a specific task, those that do should consider the data protection and information security needs within their governance plans.

Recommendations

1) Consider defining governance on the use of LLMs, including permitted use cases, what information to provide users on LLM applications used (e.g., what not to share in the prompt, and whether a model is hosted internally or externally), staff training and risk monitoring approach.
2) If own data is used to further improve the performance of LLM for task specific activities, take due care with respect to information and data protection as well as implementing mechanisms to improve the safe use of these LLMs.

6.2. Help users maximise value from LLMs

Organisations can maximise value from enabling the use of LLMs by providing or facilitating training to staff on, for instance, prompt engineering.

Organisations should acknowledge the changing nature of LLMs and consider change management activities that allow users to maximise value from LLMs. These will typically include training, preferably on-demand training, that is widely accessible in the organisation and bite-sized. It may also include setting up a support team that can be contacted to help users fine tune their most critical prompts, or building support tools for the use of LLMs, such as a tool fully controlled by the organisation to screen for personal or confidential information in prompts/inputs before these are used in the LLM.

Recommendations

3) Consider change management activities to help users maximise value from LLMs, including training, LLM support team or prompt/input pre-screening tools.

6.3. Collaborate and share experiences

Considering the fast-changing nature of AI, sharing experiences as a network is key as it reduces uncertainty, promotes a quicker common understanding and acts as a regulatory knowledge management tool that can help agencies shape their investment and experimentation.

Using fora like the European Specialised Expert Community (ESEC) and in particular the AI Special Interest Area is an effective way to share experience across the network. Another way is through sessions delivered through the EU Network training centre and to staff.

Recommendations

4) Share experiences with the use of LLMs, including on issues identified and how they were addressed, through existing fora, such as the ESEC's AI SIA, EU-NTC and others.

7. Conclusion

Guiding principles for the use of LLMs in regulatory science and for medicines regulatory activities are essential to ensure these technologies are used safely, ethically and effectively by all staff across the network.

In this fast-evolving field, it has never been more important to provide a structured approach to integrating advanced AI tools into regulatory processes to safeguard public health and maintain trust in medicines regulatory authorities.

A one-page infographic supporting these Guiding Principles on the use of LLMs is available.