Politecnico di Torino

Nexa Center for Internet & Society

Public Consultation on Competition in Generative AI

Call for contributions by the European Commission

March 2024

Studying the Internet, exploring its potential & experimenting new ideas

Name: NEXA CENTER FOR INTERNET & SOCIETY
at Department of Control and Computer Engineering (DAUIN)
at POLYTECHNIC UNIVERSITY of Turin (Italy).

Via Pier Carlo Boggio 65/A, 10129 Torino, Italia

http://nexa.polito.it

Identification number in the Trasparency Register: ID N° 234305412738-21

Type Of Respondent: Other

The Nexa Center for Internet for Internet & Society at Turin's Polytechnic University (Department of Automatic and Informatics, DAUIN), is an independent research center, focusing on interdisciplinary analysis of the Internet and of its impact on society. It was created in 2003 and has coordinated various EU thematic networks (including Communia and Lapsi). Understanding the Internet, its limitations and potential, is an indispensable course of action to ensure economic, technical, scientific, cultural and social development for the years to come; where possible, from any standpoint. For additional information see http://nexa.polito.it

Acknowledgements

Authors at Nexa Center would like to express their sincere appreciation to Dr. Carlo Blengino, Dr. Marco Ciurcina, Prof. Ciro Cattuto and Prof. Giancarlo Ruffo for their valuable insights and comments.

1) **What are the main components necessary to build, train, deploy and distribute generative AI systems;**
2) **What are the main barriers to entry and expansion for the provision, distribution or integration of Generative AI systems or components.**

Building, training, developing, distributing and maintaining Generative AI systems involve a series of complex, multidisciplinary and interconnected steps which take into account both hardware and software.

It is possible to divide the creation and distribution of such systems into three different parts:

1) The first is all about data collection and processing. All generative AI models require large amounts of data whose quality and volume significantly affect their performance. Data diversity should be a primary objective in order to avoid biases, as navigating the ethical and legal implications of data use to avoid legal consequences. This includes attention to appropriate working conditions and protection of rights for data workers, as the current industrial practices lag behind this aspect. Following this, the focus shifts to training and model development, a phase that demands substantial computational resources.
2) The second is about training and model-development. Large amounts of data are used to train AI systems through machine learning. In order to do so, a huge computational power is required. Externalities to the environment needs to be accounted.
3) The third is about distribution and monetization. After the AI system is developed, profit must be made from it. Therefore, it has to be distributed and shared to the public. The deployment strategies often leverage cloud services and APIs, with monetization models varying from subscriptions to pay-per-use schemes.

Each of these steps gives rise to critical considerations and involves specific challenges, each of which is paramount for the success of the newly born generative AI system.

Although as the field of Generative AI will progress and so will the hardware on which it runs, only a handful of players on the market are at the moment able to create a complete AI model from scratch, as the steps necessary often involve prohibitive costs, data availability, or know-how requirements.

Small and medium-sized companies are clearly on the backfoot in this new market, and some proposed solutions may cause their position to become even more precarious, such as the limiting of web scraping, which would harm small and medium enterprises, that may not have the resources to access large datasets in other ways. In a context where content production is massive, web scraping becomes an even more crucial tool to remain competitive, allowing even smaller companies to access the necessary information to train their AI models or to inform their market strategies. At the same time, those who do not adapt to the new widespread use of AI will inevitably fall behind.

Falling behind is not just a risk derived from the inability to concretely create a model from scratch. On the contrary, the risk is also due to the astonishing speed and quality of content production generated by AI, such as code or texts, which, for business enterprises, lay the foundation for a new era of efficiency. However, this also raises questions about the

expectations on the volume of work that can be produced in a given timeframe. There is a risk of creating unrealistic expectations about the amount of work an individual should be able to perform, not taking into account the uniqueness and creativity that characterize human contribution. At the same time, adopting these technologies is not just a matter of maintaining competitiveness but can also become a critical factor for survival in the market.

For these small and medium enterprises, integrating AI technologies like Copilot or ChatGPT into their workflows is often no longer an option but a necessity. Businesses that adopt these technologies early will have a significant competitive advantage over those that are slow to adapt. AI can automate repetitive tasks, accelerate the development of products and services, and provide data-driven insights that can guide business decisions.

From these considerations it can be inferred that access to advanced AI technologies could be seen as a new form of collective essential facility, where access to at least one of these technologies becomes crucial to operate competitively in many sectors. This raises important questions about regulation and fairness of access to such technologies, ensuring that small and medium companies are not excluded from the market due to insurmountable entry barriers.

There are, however, potential solutions or at least methodologies that are able to heavily assist small and medium enterprises in obtaining effective AI systems.
Firstly, **regarding 7) the role of data and its relevant characteristics for the provision of generative AI systems,** the emergence of data brokers offering clean, ready-to-use datasets can significantly impact how companies obtain and utilize data. This emergence has created an entire ecosystem of entities that facilitate the training of AI models with data that might not have been accessed otherwise, due to legal or technical constraints, raising both opportunities and ethical considerations. This new "competition inside the competition" of data brokers extends beyond the sheer volume of data they offer. Quality, relevance, consistency, and the legal compliance of data play equally critical roles in differentiating their services in the market. Legitimacy of data acquisition will remain an important topic for the foreseeable future. The landscape also suggests a shift in how companies approach data acquisition for AI training and other analytical purposes. Instead of investing heavily in collecting raw data and undergoing the arduous process of cleaning and structuring it, companies can now leverage these brokers to access high-quality data more efficiently.

Another potential scenario which is looking more and more relevant is that of data generated by AI is used to train other Ais. This trend could potentially amplify the influence of big players in the market of AI, creating an even more brutal vicious cycle that will pose another barrier to the entry of new enterprises. Indeed, large companies often have the resources, infrastructure, and data to train sophisticated AI models, which in turn can generate new data or insights that feed into further training. In this scenario, the speed of training becomes exponentially faster, due to the correspondingly higher availability of data. However, risks increase exponentially as well, and are mainly related to the amplification of biases. A dataset which originally contained biases, incorrect information or downright bad quality data can amplify its defects when used to train other models, which in turn contain vices they will spread to newer models still and so on and on.

Secondly, there already are methods through which the astonishing computational power requirements to train a Generative AI system can be avoided or at least made less important. While it is true that computational power is the definitive limit in the birth of a quality AI model, there are two phases which need to be strictly separated and defined: a training phase from scratch, and a subsequent inference phase. The training phase was already described at the beginning of this document. Following the training phase, the model enters the inference stage, which is markedly less resource-intensive. During this phase, the trained model is applied to new data to make predictions or decisions. It's at this point that the model's practical utility for small and medium companies is realized, as the same are able to acquire a pre-trained model which needs to be inferred with only a small quantity of data relevant to the specific use that the AI will serve. At this point, the computational requirements are significantly reduced compared to the training phase.

Fine-tuning represents a critical step for optimizing the AI model for specific use cases. This process involves adjusting the model's parameters by introducing new data sets, typically from the end user, to a single "layer" or a very limited part of the model. This approach allows for the customization of the model to better meet the particular needs of its application. Again, as it was the case with data broker companies, there are already many market players who offer fine-tuning services as their primary monetization strategy.

Technological advancement can further reduce the computational and economic costs of the training phase. This is especially true if the inference process can be optimized, or if techniques like pruning (which reduces model complexity) or quantization (which reduces the precision of calculations) are employed. Additionally, advancements in hardware technology have made it possible to deploy advanced AI on devices with limited computational capacities. Examples include Nvidia’s "Chat with RTX" and Stable Diffusion for generating images from text, which can be used by the end user locally on their own PCs, provided they have at least a mid-range graphics card (GPU).

Also, different types of learning are becoming more widespread, such as federated learning. This is a new approach to training machine learning models that addresses some of the most pressing issues related to privacy, data security, and the efficient use of bandwidth. Instead of centralizing data on a single server or location, federated learning allows the model to be trained directly on the users' devices. This means that the raw data collected from individual devices is not shared or transmitted; instead, only model updates or gradients are sent to a central server. The server then aggregates these updates to improve the model, which is subsequently distributed back to the users' devices. This cycle continues, with the model progressively learning and improving over time. This should, in theory, be a much more efficient training system, requiring less bandwidth and less computing power, fostering decentralized innovation while at the same time being (at least on paper) more privacy-friendly.

Generative AI is a peculiar technology in the sense that it is predominantly under the control of the industrial sector, and, as we’ve seen, primarily in the hands of large tech companies rather than small enterprises; also academia seems to play a lesser role than in other research area. From this follows that there exists a significant asymmetry between the public and private sector rarely heard of in other technologies: public sector competitors and nonprofit players

lack the expertise to compete in this space. This imbalance affects all aspects of production dynamics and marks a departure from past practices where academia played a central role in sharing innovations early on. Today, sharing within the generative AI field is almost nonexistent, with models often residing within inherently private infrastructures.

From this consideration stems the problem of **6) open source generative AI system competition with proprietary AI generative systems.**

In general, the open/closed dichotomy in AI development is not clear cut. There exists a spectrum ranging from completely closed to entirely open systems. The reproducibility of a model serves as a good indicator of openness, and is particularly important in academia for models based on scientific studies that result in publishable scientific papers, where demonstrating the model's functionality in depth is necessary for replicability. However, the current reproducibility crisis in AI research clearly shows that the current level of openness does not yet guarantee a sufficient degree of reliability for many technical and scientific advancements in the fields. This jeopardizes transferability of research results to industrial applications.

A very important line of difference lies in the transparency and accessibility of information regarding how the models are trained. Closed source AI models present a significant challenge in this regard, as their training processes from inception are not disclosed, making it impossible to replicate their functionality or understand their inner workings from an external perspective. On the contrary, some models (including some open-source models), disclose all the data used, the procedures followed from the initial training phase, and any other relevant details, thereby improving reproducibility of the system by third parties.

However, the definitions of open and closed in relation to AI are fraught with ambiguities. Meta, for instance, claims to champion open-source but faces openness problems. It does not provide access to the data, and its license is not compliant with the open source definition; yet, it is labeled as open source because it allows users certain freedoms to use the models as they see fit, within specific limits. This practice can be criticized as a form of "false open" source or “open-source washing”.

The confusion about the definition of open and closed in the AI domain complicates the issue. Actually, activities are being performed, including by the Open Source Initiative, to clarify the definition. At this stage, it is not clear if the open source AI system definition will include and consider the availability of the training datasets. It’s a matter of fact that the lack of explainability makes difficult, among other, to identify and address biases. This is why it is strongly advisable, not only that the AI systems are made available with an open source license, but also to foster the availability of the training datasets of the AI systems. This is more difficult to obtain in closed source models.

Another topic which is becoming increasingly important as AI usage spreads more and more is that of interoperability. So to answer the question **7**) **What is the role of interoperability in the provision of generative AI systems or components?** we firstly need to define what is interoperability in the world of AI.

Generally, interoperability in AI systems refers to the ability of different AI models and platforms to work together seamlessly, exchanging and making use of information across various systems and applications. This concept is crucial for creating scalable and easy-to-use AI ecosystems that can leverage the strengths of diverse systems to better deliver their outputs. It is important to notice that while interoperability can exist between AI systems, it can also exist between AI systems and normal software programs, or between non-AI software, although this last case is not taken into account here.

The potential for interoperability should theoretically vary a lot between open and closed AI models. Open-source AI models should generally offer greater potential for interoperability, because their inner workings, data formats, and communication protocols are accessible to developers. This openness would allow for easier integration with other systems and easily permit customization to meet specific interoperability needs.

On the other hand, closed systems, such as those utilizing proprietary APIs like OpenAI, may appear to be less interoperable due to their controlled access. However, if these systems are designed with well-structured, robust APIs, (such as they often are) they can, in practice, offer significant interoperability advantages. OpenAI, for instance, provides a clear and comprehensive API that enables developers to integrate advanced AI capabilities into a wide range of applications easily. The quality, documentation, and support provided for these APIs can make proprietary systems surprisingly flexible and accessible, facilitating seamless interaction with other systems.

As it has been the case in many other fields of IT, interoperability of AI systems will cause the emergence of industry standards, which would simplify a lot the work required to make different systems communicating and interoperable. In general, standards such as those we're discussing provide a common framework and set of protocols for data exchange, model communication, and system integration.