Center for expertise for
digital platform regulation

ChatGPT and the rise of conversational AI models

GPT-4, Stable Diffusion and GitHub Copilot are some of the many generative models of artificial intelligence that have emerged in recent months. Thanks to their performance and ergonomics, they are no longer reserved for experts only, but are now open to the public at large. Among them, the so-called "conversational" models have caused a lot of reactions. Capable of responding to requests formulated in natural language, they can hold a conversation with a human being, whatever the level of technicality. Their popularity is already proven (ChatGPT has over 100 million monthly active users) and their applications are numerous.

How are these conversational models built? What place do they occupy in the natural language processing (NLP) ecosystem? How are they innovative? This issue of "Shedding light on..." invites us to dive into the heart of the technology of these m in order to understand their main challenges and limitations. Combining several pre-existing technical building blocks, these AI models are not so revolutionary compared to previous NLP models, but rather innovate in terms of accessibility. Their design raises many questions about privacy, intellectual property and the openness of science: where GPT models are developed as actual products, more respectful and open models could also find their place. Educating users will be key in going beyond the hype around only a few models and truly making their capabilities accessible to all.

Shedding light on...

April
2023
#06

ONE-PAGE ESSENTIALS

Neural networks have disrupted many fields of computer research, including natural language processing (NLP). As **computer and statistical models** are trained using large volumes of data, **their predictive results depend entirely on the distribution of this data.**

NLP neural networks have taken the form of large language models (LLMs). Their training is very costly in energy (to train GPT-3, the forerunner of ChatGPT and GPT-4, the equivalent of the annual consumption of 275 French households was required) and often uses huge volumes of data from the web or copyright-free literary works. **The web contains a wide variety of data, including hate speech and sensitive content, proprietary and licensed data, and personal data. LLMs that are trained using this type of data may present compliance risks**.

Conversational LLMs such as ChatGPT and GPT-4 allow LLMs to be used by as many people as possible for more varied tasks. These models predict the most plausible text in consideration of user expectations and training data. For this reason, they inherit the biases of the human beings who were involved in their training. The data needed to train these models is sometimes tagged by workers from developing countries, who must then read large volumes of sensitive content. Despite the precautions taken by stakeholders, **misuse of conversational LLMs often remains a possibility.**

The cost of training LLMs, in terms of energy and compliant data of sufficient quality and quantity, represents a significant barrier to entry. Today's best performing LLMs are therefore often developed by companies that do not always make them openly available, due to concerns about their misuse. **With that said, some researchers are developing open LLMs.**

Conversational LLMs are likely to transform many fields, including the search engine market. In certain cases, it is more convenient for a user to read a concise answer in natural language than to refer to one or more pages, however relevant they may be. **However, conversational LLMs are currently unable to indicate which sources were used to generate an answer. They can also experience "hallucinations", fabricating plausible but false answers.**

The European regulation on artificial intelligence, which is currently under discussion, is now setting its sights on defining general-purpose artificial intelligence systems which would cover LLMs. The purpose of this regulation is to define the suppliers' transparency and compliance obligations.

NEURAL NETWORKS: THE CORE OF CONVERSATIONAL MODELS

What are neural networks?

A neural network is a model based on statistical hypotheses and rules, which is trained using large volumes of data. This training tends to mimic that of the brain: the input data supplied passes through several layers of "neurons" before providing a result. Each neuron contains **parameters**, i.e. numbers that help detect recurring patterns in input data by making simple calculations. The "error backpropagation" algorithm then tells the network whether the result is correct so that it can adjust the parameters in the neurons (see Figure 1). This operation of inference and then backpropagation is performed very many times over many examples, in order to achieve robust performance on new data. In this sense, **the model trains itself to be right as often as possible with regard to the statistical distribution of the input data**.

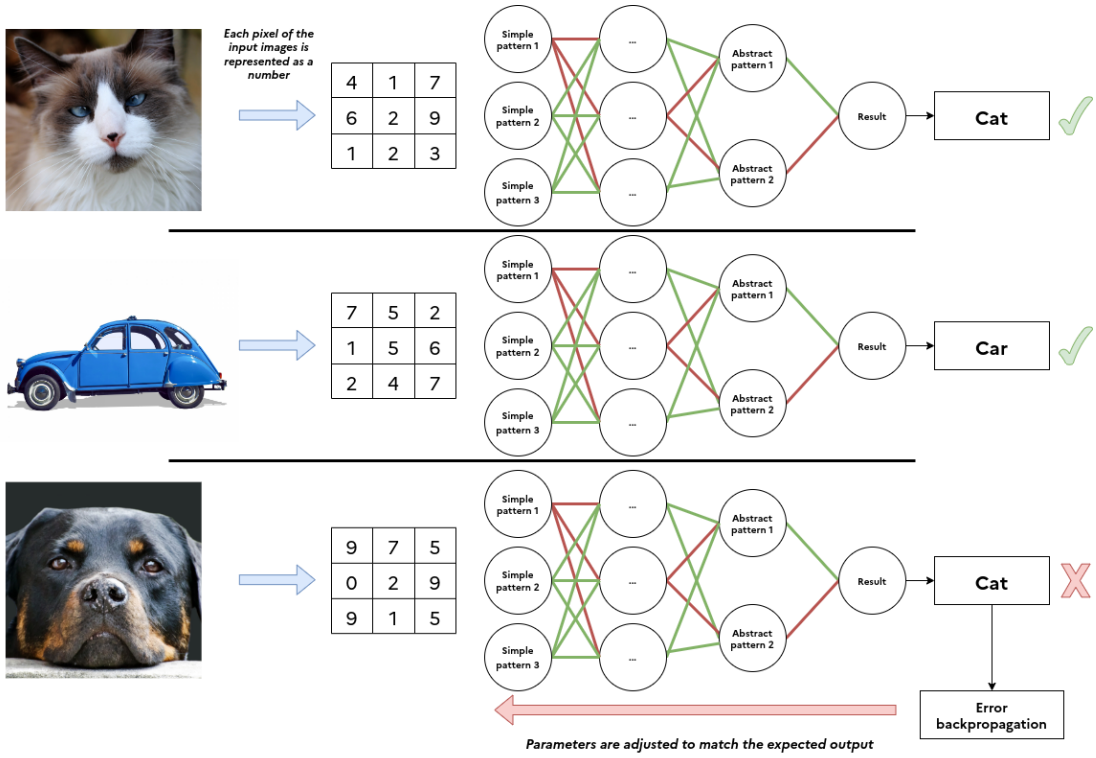

Figure 1: Simplified representation of the training of a classifier neural network.
Each layer detects recurring patterns from the previous layer. The deeper the layers are located in the network, the more they process abstract patterns. Neurons specialise by themselves: it is never specified that to differentiate a car from a cat, the presence of wheels is an indicator. It is this specialisation without human supervision that makes neural networks difficult to interpret.

The structure of a neural network is called its **architecture**. It allows an identical neural network to be reconstructed, by indicating the number of layers, the number of neurons in each layer, and finally the type of neurons. In Figure 1, the architecture consists of a first layer of three neurons, several non-detailed layers, a penultimate layer with two neurons and finally a result layer with one neuron. However, imitating the architecture of a model does not always allow comparable results to be achieved: it is the data examined during training that is the linchpin of the model's performance.

In 2012, a neural network called AlexNet overcame a major obstacle in image classification for the first time. This milestone marks the advent of the neural network revolution, with these networks producing better results than previous algorithms in many fields, including natural language processing.

Neuron networks and natural language processing

One of the first major innovations introduced by neural networks in the field of natural language processing was word embedding. This approach allows a neural network to train itself to produce a mathematical representation of words without human supervision. With this representation, mathematical operations can be performed on words:

King – Male + Woman = Queen
Paris – France + Ukraine = Kyiv

The approach consists of predicting a word from its context (such as a text with gaps, see Figure 2), or the reverse i.e. the context from a single word. The problem with this approach is that the importance of the words in the context is not measured: the model assumes that all words in the context have the same importance. For example, in the sentence "the cat chased the rat, then it ate it", the pronoun "it" could refer to the cat, but it cannot simply be inferred from the words in context (it could also refer to the rat). A model that is not able to handle this type of complex sentence is necessarily limited when it comes to translating or summarising texts in natural language.

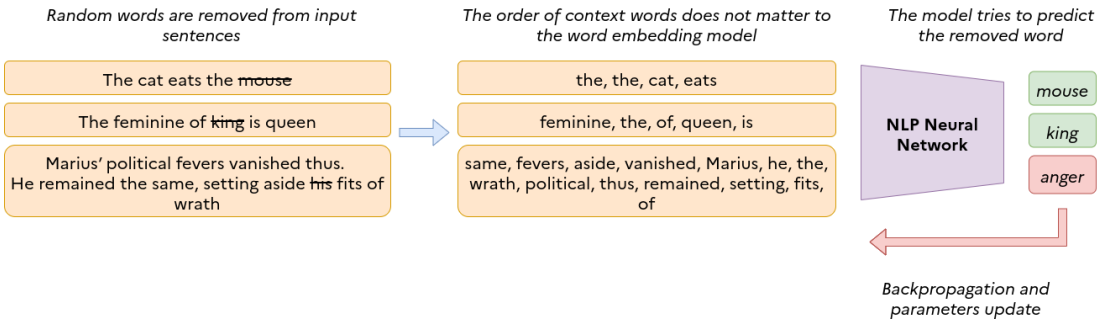

Figure 2: Simplified representation of the training method for the first word embedding neural networks. The model does not take into account the order of words in the context. It therefore struggles when faced with complex sentences.

The large language models (LLM) attention mechanism

The attention mechanism provides a solution to the problem of complex sentences. The model trains itself to recognise the importance of the words in the context based on the word examined. Going back to the previous example – "the cat chased the rat, then it ate it" – when the model considers the first instance of "it", focus will be placed on the word "cat". This mechanism, proposed by Google in 2017 in a now famous research article,[1] is the latest major advancement in the field and has spawned architectures known as Transformers (see Figures 3 and 4).

[1] https://arxiv.org/pdf/1706.03762.pdf

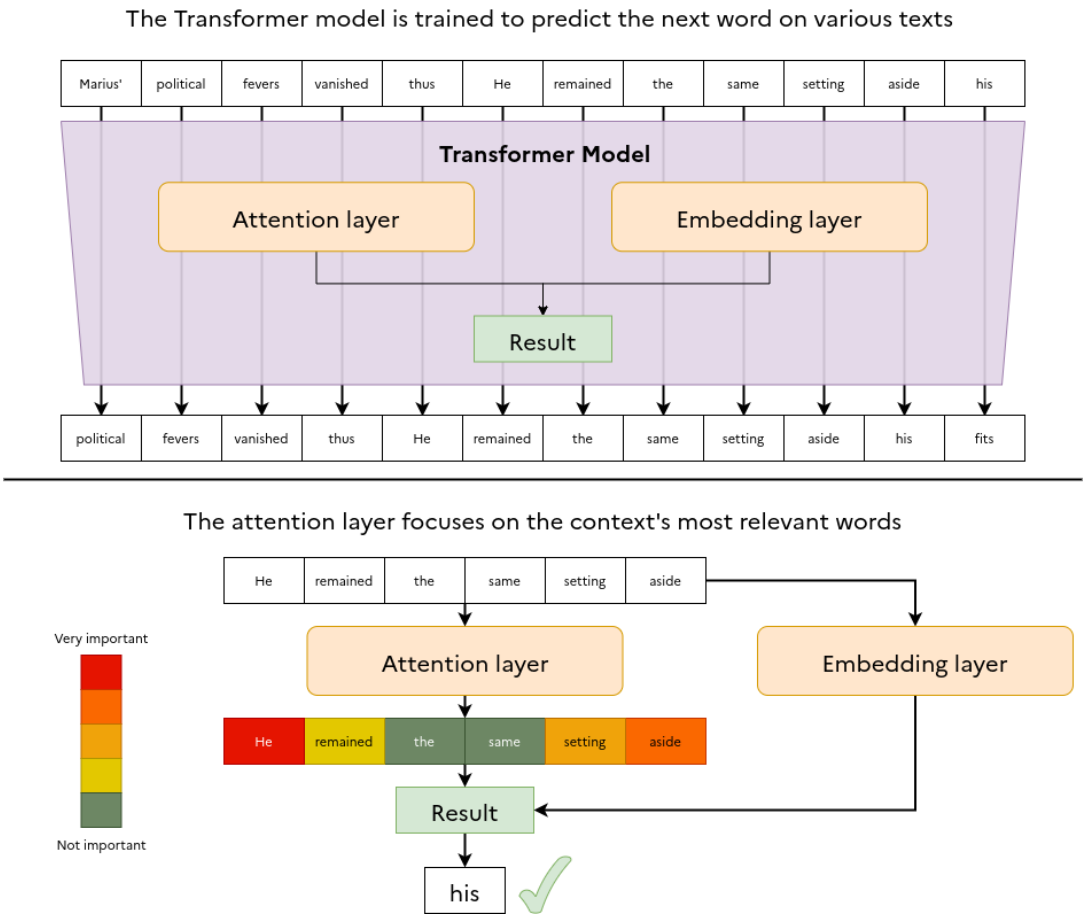

Figure 3: A simplified representation of how a Transformer model works.
The attention layer focuses on the important parts of the sentence based on the word to be predicted.

Transformer architectures quickly took the form of large language models (LLMs), albeit with major technical constraints:

- billions of parameters are required;
- a huge corpus has to be compiled;
- training just one of these models is very time consuming (several days to several weeks), hardware intensive (several dozen or even hundreds of advanced processors) and therefore energy intensive. For example, the energy required to train GPT-3, an OpenAI LLM, was estimated at 1,287 MWh,[2] equivalent to the average annual energy consumption of 275 French households.

As with word embedding, the initial training of LLMs consists of supplying the model with the beginning of a sentence or text and letting it predict the next word using, in this instance, the attention mechanism i.e. taking into account the relative importance of words in their context. This prediction operation is repeated for millions (even sometimes billions) of examples. The training data for LLMs comes from the internet (various web pages, Wikipedia in full in several languages), and from publications or textual data internal to the companies developing these models.

Like any neural network, a LLM "predicts" the most plausible result (in this case the next word) based on the statistical distribution of the training data. This ability to

[2] https://arxiv.org/pdf/2104.10350.pdf

generate plausible text based on context can be directly used for malicious purposes: many laboratories and companies refuse to publish their models in order to avoid them being used to misinform or offend.

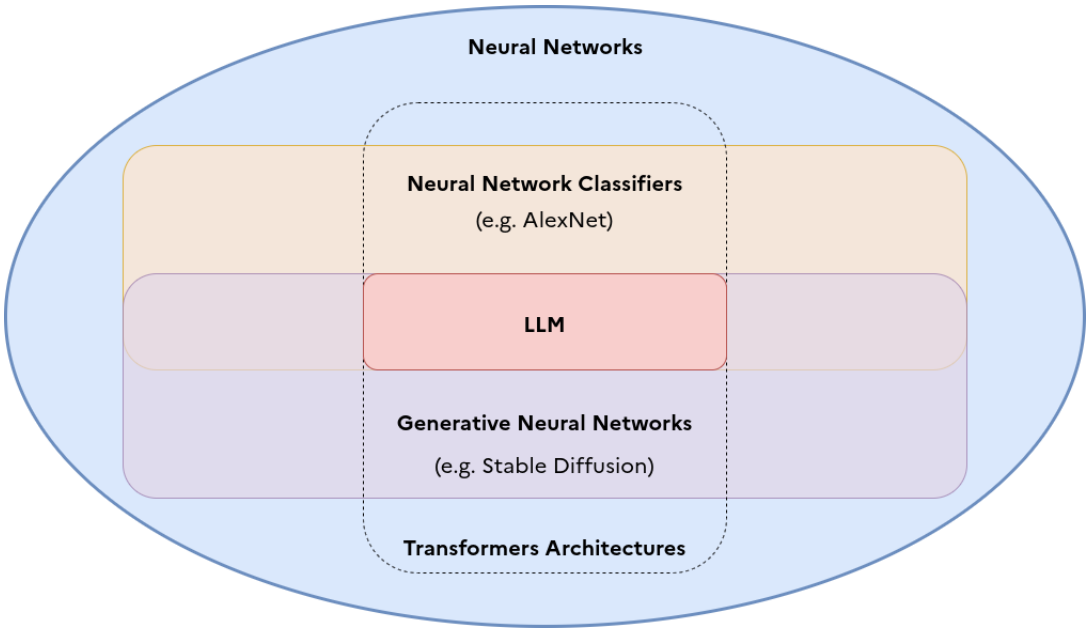

Figure 4: A schematic overview of the position of large language models (LLMs), including GPT and BLOOM models, within the neural network ecosystem. LLMs are at the intersection of Transformer architectures, classifier networks (those that classify objects provided as input into different categories, like AlexNet) and generative networks (those capable of creating content, like DALL·E or Stable Diffusion for image generation, or ChatGPT for text generation).

FROM PERFORMANCE TO ERGONOMICS: CONVERSATIONAL LLMS

Despite the apparent simplicity of the initial training task (predicting the next word using large volumes of text), LLMs can be subsequently re-trained for more specific tasks using much smaller volumes of data and achieving optimum performance. **As a result, they can accurately translate texts (including of a technical nature), summarise documents or publications, answer questions and even generate code or text**. Today, LLMs are already integrated into many tools such as machine translation and suggestions for replies in messaging or email applications. However, there are two major drawbacks preventing them from being used directly by the general public for more general tasks:

- re-training a LLM for a specific task requires some understanding of how they work;
- predicting the most plausible text is different from predicting text that meets user needs. Some questions or requests can cause LLMs problems: data from the internet or Wikipedia does not allow the model to "learn" how to respond to a user's specific needs.

To overcome these drawbacks, companies and researchers in the field have been working to develop new models: **conversational LLMs, models capable of holding a conversation with any human user.** The development methods for conversational models (or chatbots) are similar to strategies for successfully simulating a human, i.e. succeeding in the imitation game, now known as the Turing test.

In 1950, Alan Turing unveiled the first-ever method for assessing the intelligence of a machine in his research paper *Computing Machinery and Intelligence*.[3] The premise of the test can be summarised as follows: "Can a machine succeed in holding a conversation as intelligently as a human?" Turing does not suggest evaluating the consciousness of a computer program, but rather deeming a model capable of deceiving a human being intelligent. In this respect, ELIZA, the first model to pass the Turing test, developed by MIT in the 1960s,[4] simply rephrased the user's input in the form of questions, in the manner of a third-rate psychologist.

Since ELIZA, conversational models have evolved significantly. Of the most recently released ones, or those described in research articles, examples include ChatGPT,[5] LaMDA,[6] Bard[7] (the version of LaMDA that will be integrated into the Google search engine) and BlenderBot 3.[8] While each model has its own specific features, they share a technological component that has captured the general public's interest: **reinforcement learning from human feedback** (RLHF).[9] RLHF involves training a model to replicate human judgement and then using that model as a "teacher" capable of assessing other models. Figure 5 details this process.

CLOSE-UP on reinforcement learning: models that learn while playing?

Reinforcement learning consists of training a model to play a "game". Here, the game refers to both a rules-based environment and a reward linked to the model's actions. This definition of a game applies as much to genuine games, like Go or video games, as to real situations like a robot learning to walk, run, jump, etc. The model "learns" to adapt its actions to the state of the environment to maximise its reward i.e. its score. AlphaGo, the first model to beat champions at the game of Go, had been trained using reinforcement learning. Reinforcement learning from human feedback is a specific method of reinforcement learning that involves converting human preferences into a score, as shown in Figure 5.

[3] https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf
[4] https://dl.acm.org/doi/10.1145/365153.365168
[5] https://arxiv.org/pdf/2203.02155.pdf
[6] https://arxiv.org/pdf/2201.08239.pdf
[7] https://blog.google/technology/ai/bard-google-ai-search-updates/
[8] https://arxiv.org/pdf/2208.03188.pdf
[9] https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

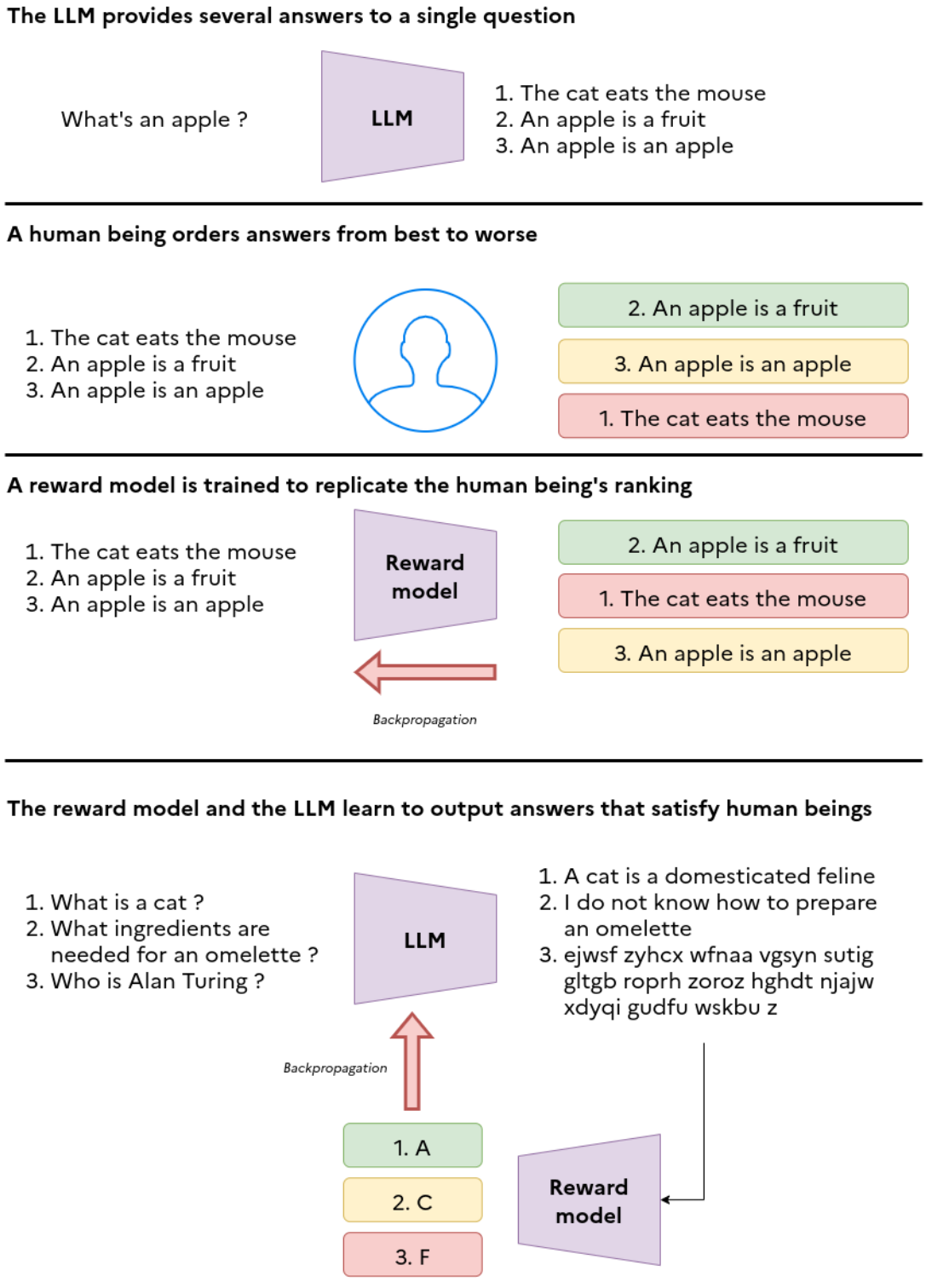

Figure 5: Use of reinforcement learning from human feedback for conversational LLM training.
The tasks of classifying the responses and of scoring them are similar: training in the former allows the latter to be successfully inferred.

In this respect, **conversational LLMs are optimised to satisfy human users insofar as possible**. This new ability allows them to hold a conversation in a very plausible manner and to take into account the needs of the users. This approach also has several limitations:

- **There is a key difference between user satisfaction and the truthfulness of the answers**. The model tends to align itself with the user's stance when the user expresses their dissatisfaction with the answer provided. In addition, conversational LLMs are trained on corpora tagged by human beings. These humans tag and evaluate their interactions against their own ideas, beliefs and stereotypes. These biases must be acknowledged: the model provides a certain vision of the world and will in turn influence users.
- Optimising user satisfaction does not restrict the generation of hate speech or illegal content. **If a user wishes to generate hate speech, the model will in theory try to comply with this request**. Conversational LLMs embed classifiers in order to detect sensitive questions or answers, restricting certain malicious applications of the model.

Despite the enthusiasm of users for conversational LLMs, their applications are in line with what had already been possible with previous LLM models in natural language processing. **Conversational LLMs generate text like any other LLM, but are better able to adapt to all kinds of users.** Innovation can therefore primarily be achieved by making these models:

- more ergonomic;
- less harmful, i.e. restricting their ability to generate false or sensitive content.

However, **software components designed to restrict the harmful capabilities of the models are all the more difficult to build because the models are efficient: the objectives of restriction but also meeting user needs are at odds with each other**. For example, a recent article about GPT-4,[10] the latest model from OpenAI, reports on an improvement in the truthfulness of the answers accompanied by a decrease in "incorrect behaviour" when faced with "sensitive" or "disallowed" content. However, data relating to these metrics or definitions of "incorrect behaviour" or "sensitive or disallowed content" is not public. Independent parties such as NewsGuard[11] have evaluated the model using their own scenarios, finding that GPT-4 is less safe than ChatGPT.

THE CRITICAL IMPORTANCE OF TRAINING DATA IN LLM DESIGN

Another key factor in the success and evaluation of an artificial intelligence model is the data used for its training. Their quality and volume are at the core of the performance of any neural network. LLMs, which use huge corpora in many languages, are an example of this. However, these huge corpora can lead to pitfalls. The training of LLMs often involves the Common Crawl,[12] a vast set of unfiltered web pages harvested from the internet, available for free. The size and variety of this dataset allows models to achieve an excellent performance level, but it may contain personal, licensed or copyrighted data.[13]

[10] https://cdn.openai.com/papers/gpt-4.pdf

[11] https://www.newsguardtech.com/fr/misinformation-monitor/march-2023/

[12] https://commoncrawl.org/

[13] The European Commission considers that the European legislation in force relating to the use of works to train AI models provides a good compromise between the protection of authors and innovation, as shown by this answer to a question asked by a Member of the European Parliament.

Several lawsuits have already been filed in the United States against companies offering generative models for intellectual property disputes.[14] LLMs could be used to generate false claims "in the style" of an author, or worse, plagiarise written works examined during training (for example a short story published by an author under certain licences that becomes part of the Common Crawl data). LLMs are also not able to reference the sources used to generate content. Finally, there is the possibility that an attacker could use these models to access personal data examined during training or shared by users during their interactions.[15]

While LLMs are often able to generate content in many languages, their accuracy and the richness of their vocabulary depend on the training data. This problem is exacerbated in the case of conversational LLMs: for example, LaMDA (Google) was trained on a corpus comprising more than 90% English-language data, and BlenderBot 3 (Meta) on a corpus solely in English. For its part, ChatGPT (OpenAI) may use pseudo-anglicisms when generating content in French (it may for example use the word "condition" to refer to a disease, whereas the correct French translation would be *maladie*) and is unable to generate content in certain languages listed as "endangered" by UNESCO.[16]

Finally, conversational LLMs require more data: on the one hand, data ordered by preference by human beings (see Figure 5); on the other hand, data tagged as sensitive or otherwise, in order to limit the model's ability to generate hate speech or illegal content. For example, OpenAI used Kenyan workers to tag the training corpora,[17] who were required to read large volumes of sensitive content. The issue of data tagging is a recurring one in machine learning, and it is not uncommon for this to be outsourced to workers from developing countries. Research papers that describe LaMDA and BlenderBot 3 also mention the use of crowdworkers. Stakeholders in this field are sometimes reluctant to offer open access to their data as the tagging operation may require a significant investment.

ARE THE STATE-OF-THE-ART LLM MODELS OPENLY AVAILABLE TO THE PUBLIC?

Conversational LLMs represent an undeniable innovation in terms of their accessibility, but state-of-the-art models are rarely openly available. While their developers often raise concerns about the harmful uses of the most powerful models,[18] the issue of their openness remains central: can we really talk about open access to the public if the most powerful LLMs are owned by a small number of stakeholders? The following examples, summarised in Table 1, illustrate these different open policies:

[14] https://www.phonandroid.com/lia-midjourney-attaquee-en-justice-aux-etats-unis.html and https://www.lemondeinformatique.fr/actualites/lire-premiere-action-judiciaire-contre-github-copilot-88522.html (in French only)

[15] A survey by the security software publisher Cyberhaven showed that many users have already shared confidential or personal data with ChatGPT: https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/. Interactions with ChatGPT are used by OpenAI to train new versions of the model. This data can be disseminated by the model or interface in the event of a data leak, as has already happened once: https://www.01net.com/actualites/chatgpt-divulgue-donnees-sensibles-utilisateurs.html (in French only).

[16] https://unesdoc.unesco.org/ark:/48223/pf0000189451

[17] https://time.com/6247678/openai-chatgpt-kenya-workers/

[18] https://www.numerama.com/tech/1307322-nous-avions-tort-en-devoilant-gpt-4-openai-dit-rejeter-lopen-source-desormais.html

- Founded in 2015 as a non-profit laboratory that aimed to open out its patents and research, OpenAI became a private company in 2019 with its flagship product being a series of GPT models. Its many versions, including the latest, GPT-4, are available via paid APIs (application programming interfaces). Various re-trainings of the model have allowed it to remain competitive against competitors, including more recent models. GPT models are the only state-of-the-art LLMs available exclusively in product form. The research paper[19] describing GPT-4, while stating the innovative aspects of the model compared to its predecessors, does not disclose any specific information about its architecture, training (data, infrastructure or method) or energy cost. In this publication, OpenAI disregards the practice – typical of this field – of expounding on the method so that the scientific community can understand how it is a technological advancement. Instead, it chooses to solely set out performance levels.
- The Google LLMs presented here, GLaM[20] and PaLM,[21] represent advances in research on the scalability of LLMs i.e. the training of even larger LLMs. They require very specific and expensive hardware architectures. PaLM is accessible via a paid API within the Google Cloud platform. Google's conversational model, LaMDA, is not directly available, but it seems to be able to successfully play the imitation game. These LLMs allow Google to offer intelligent text generation functionalities via the Google workspace for some beta testers.
- As part of its commitment to more responsible AI, Meta makes its models openly available. This is the case with OPT[22] (and its conversational counterpart BlenderBot 3), a language model with the same number of parameters as GPT-3. This also applies to LLaMA,[23] a model with fewer parameters but whose performance is comparable to that of GPT-3.
- In Europe, the BigScience collective, made up of several hundred researchers, trained LLM BLOOM[24] on the Jean-Zay supercomputer at Paris-Saclay University. This model is trained to perform the same tasks as GPT in 46 natural languages (including regional or endangered languages) and 13 programming languages. The datasets used for training are all available in open source, as is the model trained via HuggingFace.[25] The model has 175 billion parameters, as many as GPT-3.

[19] See above, page 9

[20] https://arxiv.org/pdf/2112.06905.pdf

[21] https://arxiv.org/pdf/2204.02311.pdf

[22] https://arxiv.org/pdf/2205.01068.pdf

[23] https://arxiv.org/pdf/2302.13971.pdf

[24] https://arxiv.org/pdf/2211.05100.pdf

[25] BLOOM can be downloaded here: https://huggingface.co/bigscience/bloom and can be tested here: https://huggingface.co/spaces/huggingface/bloom_demo

	Year	Maximum No. of parameters (in bn)	Public architecture	Open trained model	Open data	Conversational (RLHF)	Accessible to users (UI or API)
BigScience BLOOM	2022	175	✓	✓	✓	✗	✓
Google GLaM/PaLM	2021 / 2022	1200 / 540	✓	✗	✗	✗	✓ (paid API)
Google LaMDA/Bard	2022	137/?	✓/ ✗	✗	✗	✓	✓ (UK/US)
Meta OPT	2022	175	✓	✓	✓	✗	✗
Meta BlenderBot3	2022	175	✓	✓	✓	✓	✓ (US)
Meta LLaMA	2023	65	✓	✓	✓	✗	✗
OpenAI GPT-3	2020	175	✓	✗	✓ (1st version only)	✗	✓ (paid option)
OpenAI GPT-3.5 (InstructGPT /ChatGPT)	2022	175 / ?	✓ / ✗	✗	✗	✓	✓ (paid option)
OpenAI GPT-4	2023	?	✗	✗	✗	✓	✓ (paid API and UI)

Table 1: Examples of state-of-the-art LLMs and related open policies

CONVERSATIONAL LLMS: A PARADIGM SHIFT FOR SEARCH ENGINES

LLMs have already transformed practices in many fields, and their conversational advancements will in turn alter certain applications thereof. The search engine market appears to be one of the first sectors impacted by these developments.

The most common task for a search engine is to order external results by relevance to a user's query. Originally, search engines were more efficient the better the user knew how they worked. Nowadays, market stakeholders have developed tools that can be tailored to all kinds of users.

Conversational models offer a different solution: rather than linking a query to external content that is then ordered, the model gives a summary answer in natural language. It is easier and quicker for the user to read this summary than to consult, one by one, the results offered by a traditional search engine.

The main search engines seem to be aware of the revolutionary potential of these technologies. Microsoft and DuckDuckGo have already integrated a version of ChatGPT into their solution, while Google has started testing its Bard model. Baidu and Yandex, who also provide search engines in China and Russia respectively, have talked about the development of their own conversational LLMs and a subsequent integration into their products.

Some stakeholders, such as Yann LeCun, VP and Chief AI Scientist at Meta and an eminent figure in the machine learning scientific community, have reservations

concerning the potential of conversational LLMs to replace conventional search engines:

- these models may produce answers that are, wholly or partially, inconsistent or false;
- like the original LLMs, they are not able to identify the sources used to generate the answers.

These flaws are intrinsic to the training method used for language models. **The generated content is the most statistically plausible with regard to the texts used in training, which may contain errors or manipulated content and which form the opinions of their authors. However, it is not difficult to imagine that users of these new tools will prefer their effectiveness over their accuracy.**

REGULATING LLMS

Many experts and prominent figures are calling for a moratorium on LLMs that are more powerful than GPT-4.[26] Beyond the economic issues for the companies training these models, one may question whether such a measure would prevent the misuse already made possible thanks to existing models. It is also doubtful whether a moratorium would really allow for an assessment of existing models when no new transparency obligations are in place. In this sense, **the proposal for a European regulation on artificial intelligence systems – the AI Act**[27] **– will have to factor in the issues of sovereignty and innovation, while ensuring consumer and citizen protection.**

In the version proposed by the European Commission under discussion, the AI Act would force the suppliers of conversational LLMs to be transparent. **The proposal stipulates that a human having a conversation with one of these models should be aware of it or be able to infer its existence from the context in which it is used**. However, it would restrict the transparency obligations related to content generation to audio and video content only (deepfakes). **The text content generated by LLMs, although "[resembling] substantially genuine content", would therefore not be subject to a transparency obligation**. Finally, **this proposal does not cover generative models – a grouping that includes LLMs – in the definition of "high-risk AI systems".**[28]

In its general guidance,[29] the **Council of the European Union proposes to add a definition for "general-purpose AI systems"** to the draft regulation. This definition would include LLMs, and especially conversational LLMs. The suppliers of such systems would be subject to specific obligations, separate from the obligations applicable to high-risk systems. **The European Parliament seems to be continuing to work on this front, drawing a distinction between general-purpose AI systems built on foundation models** (i.e. any model trained on large volumes of data that can be used for a large number of tasks) and systems built on simpler models.

In the long term, this proposed regulation will regulate LLMs as products, but it is not intended to deal with user protection. These products are key to the success of

[26] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[27] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206

[28] These systems refer to artificial intelligence products that could seriously infringe fundamental rights, such as biometric identification systems.

[29] https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

conversational LLMs. While these models do not represent a significant technical innovation, their service offering has brought them swiftly to centre stage: **the simplest conversational LLMs to use, accessible using an intuitive graphical interface, seem to be the most popular.** The popularity of the models appears to be unrelated to their pure performance and more related to their design as a product. New stakeholders may therefore come to the fore, as user needs will revolve around the integration and accessibility of LLMs rather than around their theoretical performance. **Open conversational LLMs are also already beginning to emerge:[30] while their performance does not match that of proprietary models, they demonstrate that the technical expertise required to develop such models is not the preserve of a few companies.**

The main limitation to the development of competition over the most popular conversational LLMs lies in the availability of sufficient, high-quality data. In order to avoid only a small number of stakeholders being capable of training such models, data specific to the conversation task will have to be openly available, in keeping with the work of the many researchers whose open-source corpora have been used to train BLOOM and other open LLMs. OpenAI is supplied with new conversations every day at no cost thanks to ChatGPT, using them to improve its paid products. **One of the major issues ahead will be educating users about these new tools so that they can choose which model to use, and ultimately to whom they give their conversation data.**

[30] Together Research Computer, an organisation of researchers, has for example developed OpenChatKit. Like ChatGPT, exchanges with OpenChatKit are used to continue to train the model and restrict malicious applications. However, the data in this case is intended to be open.

A group of Stanford researchers has re-trained LLaMA 7B using the OpenAI paid API: the model has learned to imitate ChatGPT, reaching the same performance levels as the OpenAI model, with a much smaller number of parameters. It is also open.

Through the "Shedding light on..." series of articles, PEReN proposes elements of technical analysis on a wide variety of topics related to the regulation of online digital platforms.

Legal deposit: October 2022
ISSN (online): 2824-8201
English version: Translation Centre of the Economy and Finance Ministries

---

The Center of expertise for digital platform regulation (PEReN) is a department with national scope providing expertise and technical assistance in the fields of data processing, data science and algorithmic processes to government departments and administrative authorities involved in the regulation of digital platforms. It is also involved in exploratory and scientific data science research projects.

PEReN - 120 rue de Bercy, 75572 Paris Cedex 12 - contact.peren@finances.gouv.fr

---