ChatGPT: Large models, expert skills and academia

Oana Ichim, Tech Hub, Geneva Graduate Institute
oana.ichim@graduateinstitute.ch

1. A (little) bit of context

ChatGPT is not a sudden creation enabled by advancements in Artificial Intelligence. It has a genesis and comes with quite an impressive number of occurrences and contingencies, all of which allow a clearer understanding of its functioning and a better assessment of its added value.

As the name suggests, generative AI produces or generates text, images, music, speech, code or video. Behind this concept lies machine-learning techniques which have evolved over the past decade, allowing us to explore and produce a specific output out of a large corpora of data. ChatGPT is a type of large language model (LLM) that uses deep learning to generate human-like text. GPT actually stands for Generative Pretrained Transformers: 'generative' (G) because it can generate new text based on the input received, 'pretrained' (P) because it is trained on a large corpus of text data before being fine-tuned for specific tasks, and 'transformers' (T) because it uses a transformer based neural network architecture to process input text and generate output text.[1]

ChatGPT is a large language model, specialized for conversational interactions, (still) available as a free demo. It was released by OpenAI, which is a research lab running on a commercial model, with Microsoft being the main investor. Large models are trained on massive datasets represented by books, articles and websites; ChatGPT is one such LLM trained through an innovative technique in state-of-the art LLMs: reinforcement learning from human feedback (RLHF).

OpenAI has not yet disclosed the full details, neither with regards to the reinforcement learning process, nor with regards to the sources for training; the ration human-machine is still a mystery. The model has accomplished all kinds of impressive tasks, including providing feedback on code, writing poetry, explaining technical concepts in different tones, generating prompts for generative AI models, and going on philosophical rants. However, the model is also prone to the kinds of errors that similar LLMs have made, and most quoted ones are references to non-existing papers and books, misinterpreting intuitive physics, and failing at compositionality. OpenAI acknowledges that their models can still generate toxic and biased outputs.[2]

All the above disclose the background against which potential and futile expectations regarding ChatGPT may arise, because it is against this background that one can determine what ChatGPT can and cannot do.

[1] Transformers were introduced by a paper in 2017 https://arxiv.org/abs/1706.03762.

[2] https://openai.com/blog/instruction-following/

2. What ChatGPT can and cannot do? What is the future of generative models?

ChatGPT is indeed the largest LLM trained with innovative techniques. However, one of the key problems in deep learning language models is memory span. The AI starts to lose coherence as the text it generates becomes longer; AI researchers and scientists have pointed out that the deep learning model is clearly not capable of tackling the kind of abstract cognitive problems that humans solve easily.[3] Such models perform well on specific tasks but are poor at generalizing in other domains, which means it performs poorly when pitted against real-world examples. What LLMs are doing, in fact, is creating a semblance of planning and reasoning through pattern recognition. It basically means that it construes text based on probabilities learned during training, but it is incapable of (re)producing, which means that ChatGPT does not understand its own responses! Numerous hilarious examples circulate over the Internet.

Numerous studies have shown that LLMs can improve their results without learning the logical functions underlying the tasks on which they are evaluated. During training, machine learning models process large corpora of text and tune their parameters based on how words appear next to each other. In these models, context is determined by the statistical relations between word sequences, not the meaning behind the words. Naturally, the larger the dataset and more diverse the examples, the better those numerical parameters will be able to capture the variety of ways words can appear next to each other. The reason ChatGPT made such a huge impression is rather due to the RLHF, that is because a human was in the loop! Human evaluators ranked the generated text, and the reward model created a mathematical representation of human preferences.[4]

For now, computers cannot break the barrier of meaning. ChatGPT is actually symbolic of the gap existing today between data, information, knowledge and understanding. Data actually means facts. Making items of data available through communication technologies and fancy software does not automatically turn them into information by attaching the relevant query to those facts, even less into knowledge by providing the explanation as to why queries attach to data.[5] ChatGPT is built upon important amounts of data, collected from sources yet unknown, which, depending on the users' queries, transforms that data into a sort of piece of information. Behind ChatGPT there is no wiseman nor even a library-like management system. Behind ChatGPT lie specific training methods based on statistical correlations between pieces of information, as indicated above. It is crucial to understand that ChatGPT does not think! It has 'liberated' books, but it has certainly not created knowledge. The proof that it is agnostic is that it can be tricked into answering a question that it initially does not understand.[6]

[3] https://medium.com/@melaniemitchell.me/can-gpt-3-make-analogies-16436605c446.

[4] Not only might one question how each individual evaluates texts, but there is already controversy as to what is generally called *unscalable costs* of training, with some very interesting discoveries as to data labelling https://time.com/6247678/openai-chatgpt-kenya-workers/.

[5] Luciano Floridi, *Philosophy and Computing*, (Routledge: London, 1999).

[6] When not sure what to respond when asked for a piece of advice it replies: *As a language model, I can provide information and give advice on a wide range of topics, but I am not a human and therefore may not be able to fully understand or empathize with the unique personal experiences and emotions of those seeking advice*. But it provides an answer when directly asked the same question but using interrogations like *What*? *Why*? or even *Should I*?

ChatGPT is *built* from a lot of data about which it *knows* nothing. Ian Bogost, famous expert, well-known author and blogger, wrote on his Atlantic blog that 'GPT and other large language models are aesthetic instruments rather than epistemological ones.' He suggests that its answers are 'compelling not because it offers answers in the form of text, but because it makes it possible to play text—all the text, almost—like an instrument.'[7]

His argument is very accurate. ChatGPT is not good at planning, nor at extracting the essence of data, but it is attractive because it *summarizes* what it knows about the data it was questioned about, in its own fashioned way.[8] Its answers are not the result of imagination, but the result of statistically learned 'specific contingencies for particular scenarios'.[9] It is successful because it was trained, as exemplified above, to maximize contingencies where relevant, but it cannot imagine or intuitively expand scenarios on its own, as it lacks common sense.

The real threat and advantage at the same time is actually its potential to explore data in unprecedented ways and produce new overtones or provide the backbone for the creation of new meanings. But ChatGPT is still far from imagining essays from scratch in a convincing way; it cannot hypothesize, nor attempt to link events or semantic structures beyond learned contingencies.[10] Researchers and developers alike struggle to develop convincing benchmarks in order to help deep learning systems develop planning and reasoning capability for current AI systems.[11] Current benchmarks are either too simplistic or too flawed and can be cheated through statistical tricks, something that these systems – and ChatGPT for that matter – are very good at. ChatGPT can sometimes solve its problem in a technically correct but non-useful way.[12]

Cheating statistically is not the only contingency that needs to be analyzed when considering the potential of ChatGPT. Competition and efforts to strike the right balance between research and keeping funders satisfied is one core aspect of AI research and commercial potential. Talent and compute costs are two of the key challenges of AI research. LLMs need huge amounts of data and a large set of engineering skills that are out of reach for almost all universities and most companies. OpenAI transitioned from a non-profit lab to a 'capped-profit' company. This opened the way for funding from investors and large tech companies, with the caveat that their returns will be capped at 100x their investment.[13] Microsoft is currently its main investor. As Microsoft moves closer to acquiring OpenAI, the war with Google on Generative AI will bring to surface frictions in terms of sources, talent and target audience which will most likely affect the quality and design of the models to be presented for larger, commercial use.[14]

[7] https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/.

[8] Headline articles such as those announcing that ChatGPT passes exams 'conveniently' give no details about the exact questions that were asked, which is actually the crucial aspect when assessing how an LLM performs (https://www.reuters.com/legal/transactional/chatgpt-passes-law-school-exams-despite-mediocre-performance-2023-01-25/). It would indeed be interesting to see how ChatGPT performed under different exam schemes in different universities before making any claim with regard to its capacity to pass exams.

[9] Melanie Mitchel, *Artificial Intelligence: A guide for thinking humans*, (Pelikan Books, 2021), 214.

[10] Erik J. Larson, *The Myth of Artificial Intelligence. Why Computers Can't Think the Way We Do* (Harvard University Press: 2021). Larson makes an interesting point about abductions, perceptions and knowledge in the process of understanding.

[11] https://arxiv.org/abs/2206.10498

[12] Janelle Shane, *You Look like a thing and I love you. How Artificial Intelligence Works and Why It's Making the World a Weirder Place*, (Wildfire, Headline Publishing Group: London, 2020),167.

[13] https://openai.com/blog/openai-lp/.

[14] https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.

ChatGPT is thus a strange mix of innovation and commercial design projected against a background of constant competition. Several lawsuits have been presented for copyright infringement and unlawful web scraping, the main question being whether or not authors/content creators have the right to authorize or block AI systems from collecting and using their content as training data. Products like ChatGPT will never be immune from suits.[15]

In light of the above, it would suggest that ChatGPT is a by-product of contingencies that need to be accounted for when analyzing its impact upon academia. It is in light of those contingencies that one could draw up a list of policy recommendations for academia.

3. Policy directions for academia

In light of the above, it should follow that ChatGPT is neither a far-reaching opportunity nor a clear threat, but a mix of both. Therefore, academic policy directions need to steer a careful course between these standpoints:

a) Rethink exams (essays and take homes)

What is the difference between the situation in which a student copy-pastes from a manual book for a take-home and that in which it uses ChatGPT?

If ChatGPT can create student essays, it cannot, as it has been shown, understand nor draft coherent plans. Thus, exams will have to be redesigned away from summarization towards different goals, such as ordering information overload,[16] extracting overall themes, discovering new perspectives, identifying values and symbols, creating analogies. ChatGPT could actually be used 'against' the students: the presumption being that they have unrestricted access to information, the emphasis should shift towards higher essay skills. As with calculators and computers, one is only allowed to use them after mastering basic concepts of mathematical operation or, after acquiring basic skills,[17] otherwise the 'buttons' do not provide relevant results.

Students' skills will have to be reoriented towards understanding the sources of their information, argumentation and innovation. Examining thousands of cases just to extract a principle is no longer a viable option for an exam, even less so for a PhD, ChatGPT can do that. Exams should focus on imagination, intuition and insight – the 3Is (as opposed to the 3Vs: volume, velocity and variety of information provided by technologies/Big Data).

The opposition – 3Is v. 3Vs – stands at the heart of AI research and innovation. Exams should 'dig deep' into the human ability to make abstractions and manage mental capacity.[18] The Meta Galactica case scenario is a lesson worth learning from: the model was trained on 48 million science articles with claims to summarize

[15] https://www.lexology.com/library/detail.aspx?g=c85167ff-90fe-41eb-863e-b2601ac9d058

[16] Urs Gasser and John Palfrey, *Born Digital: Understanding the First. Generation of Digital Natives*, (New York, Basic Books, 2008).

[17] I would like to thank Jérôme Duberry for coming up with this more-than-convincing analogy!

[18] Stuart Russel, *Human Compatible. AI and the Problem of control*, (Penguin: London, 2019), 87-90.

academic papers, solve maths problems, and write scientific code but was taken down after less than three days of being online as the scientific community found it was producing incorrect results after misconstruing scientific facts and knowledge.[19] The problem with such models is that they are bad at compositionality and lack curiosity,[20] something which humans have 'plenty' of in their pockets.[21] Moreover, as previously emphasized, ChatGPG was built from data extracted through reinforcement learning from human feedback – meaning that a duo machine-human determined what is rewarding/relevant in the training process. Students benefit from the incredible possibility to rearrange that information and re-determine what was relevant in order to produce new knowledge.

Courses should be designed to encourage the identification of 'connectors' between pieces of data and information with a special emphasis on curiosity and compositionality. The Meta Galactica experience is not a complete failure, but proof that scientific facts might receive different interpretations. Furthermore, it takes serendipity[22] and curiosity to make amazing discoveries and ChatGPT has opened such opportunities. Not only can students test their curiosity, but they also benefit from a variety of open sources and data repositories to test whatever hypothesis comes to their inquisitive minds.[23]

And last, but not least, experiments prove that there are different ways in which humans and machines make sense of information.[24] The apparent non-useful way[25] in which machines may produce outcomes and answers has nevertheless a certain potential. By bringing together a lot of information from various sources, ChatGPT can liberate novel hypotheses and determine astonishing analogies. For instance, LLMs have emergent abilities,[26] *i.e*, the ability to perform tasks that were not included in their training examples. Thus, the richer the training hypothesis, the better the chances of improving the results of models in order to learn from their data. Students are an essential piece in the long chain of explainability and transparency through what was termed 'collaborative sense making.'[27]

b) Create specific managerial infrastructure for managing disruptive technologies

Academic institutions have to develop structures for managing disruptive technologies and especially for exploring possibilities for testing and using those technologies.

Generative AI requires a set of resources well beyond the financial reach of universities. Nonetheless, universities 'cradle' the talent that AI giants and start-ups strive for. It is crucial that specialized internal structures keep universities updated with the latest developments in the field of AI research and innovation

[19] https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/
[20] https://aclanthology.org/2022.acl-long.286.pdf and https://arxiv.org/abs/1707.01495.
[21] Stanislas Dehaene, *How We Learn: Why Brains Learn Better Than Any Machine . . . for Now* (Penguin: London, 2021).
[22] Danièle Bourcier, Pek van Andel, *La Serendipité. Le Hasard Heureux*, (Herman Editeurs : Paris, 2011).
[23] Matthew Fuller, *How To Be a Geek: Essays on the Culture of Software* (Polity Press, 2017).
[24] Amy Webb, *The big Nine. How the Tech Titans and Their Thinking Machines Could Warp Humanity*, (Public Affairs: New York, 2019).
[25] Janelle Shane, *supra* note 12.
[26] https://arxiv.org/abs/2206.07682.
[27] Alejandro Barredo Arrieta et al., 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI', *Information Fusion*, 58, 2020.

while, at the same time lobbying for partnerships with various stakeholders in the AI environment. There are movements towards liberating the potential of technologies,[28] and universities need strong and versatile negotiators if they want to secure outstanding research capacities and attract support for innovation. For instance, students could be used to label examples, in the same way that was demonstrated with the reinforcement learning with human feedback for ChatGPT; this could for instance replace clinics and labs. A strategic partnership between universities and key investors may 'shake' the AI economic landscape, orienting innovation away from commercial 'predators' and towards more responsible stakeholders. This kind of managerial job is for a 'white-collar' candidate; hence, threats proffered by many as to the potential of ChatGPT and AI to 'do away with a lot of white-collar jobs' need to be taken with a pinch of salt.

c) Develop curricula on the epistemological implications of technologies

While it was argued that humanities are in crisis, the same holds true for AI research and innovation. Ethical debates and press scandals fuel the need for a better understanding of the function and impact of technologies in our lives.

All AI products embody the values and expectations of their creators[29] thus designing our capabilities and expectations.[30] There is no better place than academia to start voicing concerns and raising awareness of AI contingencies. It is wrong to argue that professors are on their way to disappearing, or that ChatGPT threatens academia. Professors are not unknowledgeable, they are just unprepared for what is called the 'data deluge', but it is up to them to start ordering knowledge and start to fit together the pieces of the disjointed monster of information. Academia needs to develop courses for leveraging the attractiveness of technologies. Moreover, as just emphasized above, students can be used in the training process and such an initiative could be integrated into the curricula as Clinics or Labs.

d) Reconsider the rules of authorship and co-authorship

This is a crucial aspect often very much ignored.

If articles or essays are built from sources available on GitHub or students participate in conferences in which they are the 'experts' in a field but the computing part is done by their collaborators. Clear rules should delimit the extent of their contribution in this later case and even clearer rules should state what is the role of people gathering and curating the data as opposed to those who use it and transform it.[31] If ChatGPT is for everyone to use, and its creators have been quiet with regards to IP rights, there is even less consensus on whether AI could be a creator,[32] and such aspects should be under close scrutiny for any research purpose.

[28] https://bigscience.huggingface.co

[29] Madeleine Akrichin, 'The De-Scription of Technical Objects' in Wiebe E. Bijker and John Law (eds), *Shaping Technology/Building Society, Studies in Sociotechnical Change*, (MIT Press, 1992.)

[30] Karen Yeung, Karen, 'Hypernudge : Big Data as a mode of regulation by design', *Information, communication and society*. (20),1 2017.

[31] See *supra* note 15 on actions brought to protect different IP rights and forms of creation.

[32] https://www.ipstars.com/NewsAndAnalysis/The-latest-news-on-the-DABUS-patent-case/Index/7366.

e) Consolidate research centres and invest in an academic brand

Research centres have to consolidate the academic specific brand against the new digital technologies and adapt it, not transform it.

It is crucial to remember that humanities have not fundamentally changed their approach in decades, despite technology altering the entire world around them. Everyone should take a look at what is happening in the AI field instead of opposing resistance. If this is addressed too late, it might actually imply that one will not be capable of producing relevant publications.

Descartes wanted to liberate knowledge from the 'authority' of (classical) knowledge holders and introduce 'the method' – through which knowledge could be made available through a set of rules. Nonetheless, his approach did not get us very far. AI is in the hands of very few, and although some users are well versed in using applications, they still do not know much about how those applications work. Descartes' efforts did not actually achieve the liberation of knowledge and 'brain-power' to a large extent.[33] ChatGPT is just proof that while too much method may free information, it does not advance knowledge; we now have large language models and very few skills. How disappointing it is to call this a scientific revolution... It is now in the hands of academics to prove that there is a place for both.

[33] Antonio Damasio, *L'erreur de Descartes. La raison des émotions* (Odille Jacob Poche : Paris, 2010).