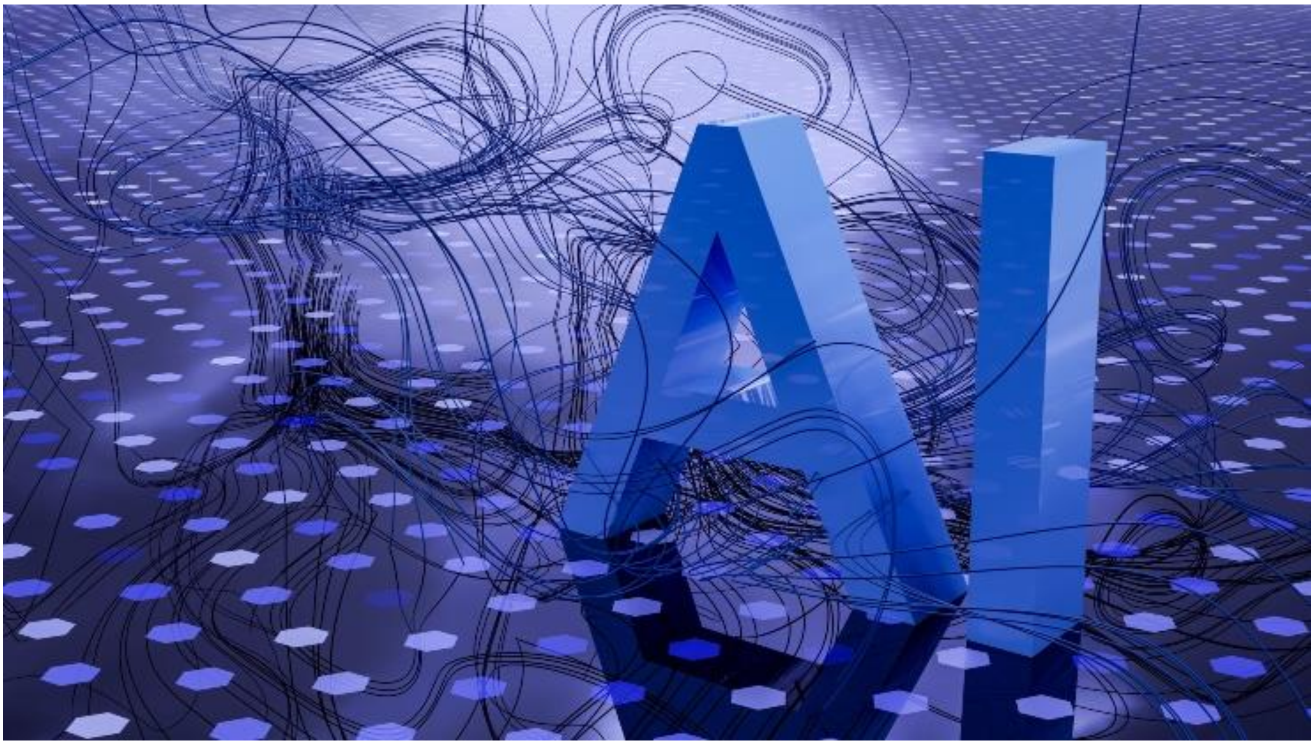
How generative AI can support inclusion



Large language models (LLMs) such as ChatGPT are prone to making things up, getting the facts wrong and infringing customers' data privacy. But they also offer many benefits and domain specific LLMs are being integrated into specialised applications in many areas. ***Christine Chow, Charlotte Bourquin*** *and* ***Enkhzul Stricker*** *find at least two use cases that can leverage ChatGPT's distinctive advantages from an inclusion perspective: therapy assistance and policy communication.*

Q: What year is it?

ChatGPT: The year is 1400.

Q: Why do you think it is 1400?

ChatGPT: Because I am a medieval AI system.

I only know about the events and culture of the 14th century.

(Source: Bubeck et al 2023: 61)

ChatGPT derives its name from generative pre-trained transformer, a chatbot or machine learning algorithm that could mimic human language interactions. It is a large language model (LLM) that uses deep generative learning techniques (which means that they generate new content) to produce human-like text through autoregression. However, domain specific LLMs are already being integrated into specialised applications in fields like clinical medicine, such as BioMedLM, formerly known as PubMedGPT.

Here we explore alternative use cases beyond specific tasks such as summarising meeting notes, managing supply chain, personalised marketing and coding. We aim to maximise the business opportunities of using this technology, whilst acknowledging the risks they present (Bubeck et al, 2023).

Challenges

Let’s start with the challenges presented by ChatGPT.

1. Hallucinations

Hallucinations mean ‘making things up’, and it is the essence of the ‘generative’ capability as labelled in GPT. Closed-domain hallucinations are errors made using only the information provided. There are opportunities for checking consistency or alignment as they are usually fact based. Open domain hallucinations are more challenging as they are often connected to creative activities that, by nature, have less constraints during the process of ‘generation’.

2. Inferential capabilities

Inferential analysis refers to understanding that is deduced from knowledge gained in a different setting. For example, in this TED talk by Dr Yejin Choi, ChatGPT is not able to conclude that it would take the same number of hours to dry five pieces of clothing compared to thirty pieces if they are put under the sun because it lacks contextual knowledge of the world, or what we deemed ‘common sense’.

Identifying errors and taking on feedback. ChatGPT can make arithmetic mistakes, but it is able to provide consistent explanation if carefully led through every step by the curator or user of the model. Human handholding is needed to correct the error.

3. Customers’ data privacy and personal identifiable information (PII)

This could potentially be addressed by using a private ChatGPT service or overlay which means that customer or personal information is not shared with, or is shielded from, the OpenAI platform. However, it is still debatable whether retaining data and computation on-device or data anonymisation are sufficient to meet privacy expectations because model parameters exchanged between customer and OpenAI that conceal sensitive information can still be exploited in privacy attacks.

Given the above challenges, we find at least two use cases that can leverage ChatGPT's distinctive advantages, specifically from an inclusion perspective: therapy assistance and policy communication.

Therapy assistance (be bold)

Can ChatGPT be therapy assistance for those with alexithymia, the inability to recognise one's own emotions and express them in words.

The rapid development of AI is fuelling discussion about its possible uses in treating mental illness or any other major depressive disorder that will require the therapist's support. ChatGPT could potentially help a person with patterns of alexithymia by providing a safe and controlled environment for communication and social interaction.

Using a tool like ChatGPT allows an individual with autism spectrum disorders (ASD) to read social interactions and emotions, as the user has the option to customise the language and communication settings in a specially designed tool for curating emotional responses. Information and communication technologies can be used in different ways and settings and can provide consistent and predictable responses that ultimately reduce anxiety and stress for the user.

Since the start of the COVID pandemic, the global demand for anxiety disorder treatments has increased. Even in jurisdictions where access to national health services is in theory making such treatments accessible, waiting time has lengthened, up to a few years, forcing patients to go private. Those who cannot afford it may have to abandon consultation and forgo the opportunity of receiving treatment. In addition, in some parts of the world, mental services are limited and shrouded in stigma. Using a platform such as ChatGPT for therapeutic assistance could save money and time while avoiding any stigma associated with mental disorders. Communicating with a language model designed to 'chat' could also reduce the anxiety of being judged when discussing a range of deeply personal issues.

Some may question whether a computer programme can mimic empathy, identify early signs of relapse – such as depression or anxiety – and distinguish the warning sign that could threaten the user. There are also concerns around bias, unpredictability, inaccuracy or even disturbing responses, instead of consistent, predictable and targeted emotional release for users as ChatGPT was trained by data on the internet.

Fortunately, it is possible to customise or finetune ChatGPT for specific use cases without retraining the language model from the start. For example, by setting specific learning objectives for therapeutic assistance or dialogue with users, subject to human therapist oversight, the arrangement can partially alleviate the workload of health services shortening waiting time, improving access to health services.

To design a therapy assistance tool that addresses the four identified ChatGPT challenges highlighted above, we put forward the following for consideration:

A. What is the 'co-pilot' arrangement for a professional therapist who uses ChatGPT as a therapy assistance tool? This is critically important to ensure any hallucinations, factual errors and problems that may negatively impact the user are addressed in a timely manner with human feedback.

B. How would the therapist use the tool to more effectively address emotional challenges or bottlenecks that negatively impacts the user?

C. How will the inferential capability advantages of human be combined with the integrative capability of ChatGPT? Is there a clear business model built on the strengths of the co-pilot combination?

D. How will the chat history be stored and used for additional training to improve the therapy assistance tool without compromising personal data privacy and security?

> *"New AI tools such as ChatGPT offer a wealth of opportunities for aiding those with a range of needs, including those with difficulties in everyday social interactions. In fact, AI has already been used as an effective therapy tool in use cases for veterans, where those returning from war are more likely to disclose symptoms to an AI therapist than to a real one. AI tools present great promise, when used carefully, to open more avenues for those seeking support for a range of social issues."*
>
> - Dr. Nikki Sullivan, Assistant Professor of Marketing, Department of Management, The London School of Economics and Political Science

Policy communication (be clear)

Can ChatGPT be a communication tool to improve the quality of policy communication?

Policy documents are notoriously complex and laden with technical jargon and intricate details. These complex policy documents can be challenging to comprehend for a broad audience. ChatGPT's ability to generate concise and accurate summaries in plain language holds promise for making policy communication more accessible and inclusive.

Hansen and Kazinnik (2023) found in their study that ChatGPT demonstrates a strong performance in classifying Fedspeak sentences. The tool was able to justify its classifications and its 'reasoning' is like that of a human reviewer. The findings suggest that ChatGPT can help analyse human feedback of policy announcements and identify areas of improvement for policymakers. This evaluation process allows policymakers to learn from past experiences and adopt more effective communication techniques.

More generally, it is possible to include ChatGPT in the workflow of producing policy communication content. It could support the drafting of well-articulated policy documents and automate part of the processes, such as summarising content from human subject matter experts, and fine tune content that is overly technical. Yokosuka City in Japan's Kanagawa Prefecture is set to pioneer the use of AI chatbot ChatGPT in public administration. Over a month-long trial, four thousand municipal employees utilise the AI tool for tasks such as summarising sentences, checking for spelling errors, drafting documents and developing copies for marketing and communication. The city's strategy is to employ ChatGPT for automating routine tasks, thereby redirecting human resources towards roles that demand interpersonal interaction (Japan Times, 20 April 2023). This trial period could potentially usher in a new era of administrative efficiency in Yokosuka and beyond.

However, a significant concern is that ChatGPT's lacks contextual understanding, as outlined in the earlier part of this article. It might not fully grasp the implications of certain policy decisions or fail to consider broader socio-political context in which the policy operates. Human oversight is essential to ensure that the information conveyed is accurate, appropriate, and in line with the intended policy objectives. This way, it can make the most of the opportunities offered by these technologies, while managing the risks they pose.

♣♣♣

Notes:

- *This blog post represents the views of its author(s), not the position of LSE Business Review or the London School of Economics.*
- *Featured image by Steve Johnson on Unsplash*

Christine Chow

Christine Chow is managing director of Credit Suisse, where she heads the area of active ownership, covering global markets and all asset classes. She is an emeritus governor of LSE and sits on the advisory board of the School's The Inclusion Initiative.

Charlotte Bourquin

Charlotte Bourquin is an analyst in active ownership and sustainable investing at Credit Suisse Asset Management.

Enkhzul Stricker

Enkhzul Stricker is an analyst in active ownership and sustainable investing at Credit Suisse Asset Management.