From AI to AGI in 2025?

Large language models, such as ChatGPT, have received a lot of attention in recent years, but there will probably be many other kinds of building blocks in future AI systems. The world of AI is still new, and there are still many new things to explore.

Notice: This english version of the article was ***automatically*** translated by *Google translate* from the original Danish version to English in December 2024.

It won't be long before all phones will be equipped with a multimodal AI program that, based on input from the phone's camera and microphone, will answer questions and comment on the world situation, what it sees, in English, Uzbek and Danish. The next step could very well be to have such models running inside humanoid robots that walk around on two legs among us. In Nancy, France, this fall, at the conference "Humanoids 2024" you could see many of the robot models that in the coming years, at least according to the manufacturers, will come out and help in warehouses, guide us in shopping malls, assist in the healthcare sector, and create security in city centers by patrolling the streets at night. First a few robots, then thousands, and later millions of robots[1].

From the self-driving cars, however, we know that it can be a long way from a vision to an AI system that can actually solve a given task on a human level. In December, General Motors withdrew from the robot taxi company Cruise. After investing more than $10 billion in Cruise since 2016, and with the prospect of having to invest even more to make a business out of it, they apparently felt that it was not that easy to go from assistance systems, helping a human driver, and then to having fully automated taxis on all the world's roads[2]. There's a difference between driving straight on a highway in California and navigating traffic in New Delhi.

That said, in recent years we've had almost daily reports of new AI systems that are well on their way to achieving Artificial General Intelligence. General intelligence that can understand or learn any intellectual task that a human can solve. In December, OpenAI[3] struck again, announcing that their new "reasoning" models have achieved a score of 87.5% on a type of pattern recognition task[4], which surpasses a typical human score of 85%. Here we are no longer "just" talking about large language models that, trained on huge amounts of text, can give the most likely answer (text) to a given question, in a given context. In these latest "reasoning" models, various techniques are used to examine the model's internal dialogue and plan the best answer before this answer is sent to the user[5].

Are we on the verge of the next big AI breakthrough?

In a now familiar pattern, not many days passed from the OpenAI presentation until the American psychologist and AI researcher Gary Marcus once again felt the need to tell the world that such generative AI systems are good at giving answers that sound good, or plausible, in a given context, but not at understanding, on a deeper level, what they are actually saying. They cannot fact-check themselves. Which then gives rise to endless so-called "hallucinations", where the system claims something that is not true. Without even fully understanding whether a given answer is right or wrong. According to Marcus, you can make good demos with such generative AI systems, and perhaps use them for brainstorming, but these systems are not particularly good components in reliable products[6]. Marcus also did not neglect to tell us that OpenAI is set to have an operating loss in 2024 of 5 billion dollars, which in his eyes makes the valuation of the company, 80 billion dollars, far too high[7].

In short, it is quite understandable that consumers and investors at the beginning of 2025 may feel a little confused about what to believe. Are we facing a breakthrough with AI that will become AGI, artificial general intelligence. Or are we exposed to "corporate hype" that exaggerates what they

have, in the hope of continued investments in products that do not quite live up to their promises, and do not generate profits that justify the investments?

The IT industry has of course been here before. There have previously been periods of great progress followed by declines in AI, the so-called AI winters. In 1973, the so-called Lighthill report was published in England, commissioned by the English parliament, in which James Lighthill concluded that AI "had completely failed to achieve its grandiose goals", and "could only be used to solve highly simplified toy problems"[8]. A subsequent decline in investments in the UK and the USA resulted in fewer AI projects and thus a so-called AI winter. In the 1980s, there was great interest in the so-called "knowledge-based" systems, which, however, also proved to have difficulty living up to inflated expectations, which in turn triggered an AI winter with a lack of willingness to invest in AI systems. For those who may have forgotten, it can be mentioned that the very word "AI" in the 1990s, and at the beginning of the new millennium, had such a controversial reputation that many of us who were educated in IT and AI during that period were told, by friendly mentors, that in job applications you should probably write that you had worked with things like Bayesian networks, control systems or similar things that sounded more "technical". It was recommended that fluffy words like AI was only used sparingly in applications.

Things really changed when AlexNet, a neural network designed by Alex Krizhevsky in collaboration with Ilya Sutskever and (2024) Nobel laureate Geffrey Hinton, proved to be the best solution for categorizing images in the so-called ImageNet Challenge in 2012. And from there, as everyone knows, the rest is history, with one major AI breakthrough after another, including the release of ChatGPT in November 2022[9]. As Wired Magazine editor Kevin Kelly wrote back in 2014: “It’s easy to predict what the business plan for the next 10,000 start-ups are: Take X and add AI”[10].

Making the large language models even better.

That said, it is of course still interesting to see whether the large language models, such as ChatGPT, which have been the subject of much of the media coverage of AI (since November 2022), are really the technique that not only provides a good solution within an area, but also an important step on the road towards general artificial intelligence, AGI, comparable to human intelligence.

Before 2022, you could easily come across language models that gave such weird answers that you did not suspect them of being particularly intelligent in any way. The models were small and not trained on particularly large amounts of data. And humans had not, after training, humans added "blocks" of unfortunate, incorrect answers. Today, you can recreate the “early-2022” experience by, for example, running the model "llama2-uncensored" on your own laptop[11]. During my testing this winter, I managed to get into a conversation, where the model claimed to have children, Jack and Alex, both of whom were humans... Things started to get even stranger, when I wanted to know more about how language models can have human children. And was told that it was a transformation, similar to the one you see when a dragon turns into a human.

The reason is of course that here you are talking to the "raw" language model. That uses "next-word prediction", where the answers is what most often follows in the texts they are trained on, with a certain probability. So this particular model had probably been trained on a text from the internet about dragons. Without fully understanding that we actually don't know that many examples of real dragons turning into humans.

ChatGPT is of course much more advanced. And is constantly improving. Yet in the spring of 2024, it was relatively easy to find examples of questions, where it didn't came up with particularly good answers.

For example, in the spring, you could ask ChatGPT: "Imagine that you are standing in front of 3 boxes of fruit. All the boxes have a label on them with what kind of fruit is in the box. One box contains apples, another pears, and the last box contains mixed fruits. All the labels are wrong with

regard to what is actually in the box. Without looking, how can you take a fruit from a box and then put all the labels correctly?". The correct answer is that you should take a fruit from the box of mixed fruits. If it was an apple, that box should say apples. The label that said apples must also have been wrong. If you replace it with the label mixed fruits, the label pears will still be incorrect, so the apple label should probably be replaced with the label pears. Etc.

When ChatGPT routinely answered this question incorrectly in the spring, it was of course a great encouragement to everyone who was not quite ready for AGI. After the summer holidays, however, the story was different.

ChatGPT was updated and could now suddenly answer this, and many other questions that it had previously been unable to answer. Many were baffled. As they were in 2023, when the public was told that ChatGPT had achieved a passing score on certain American law courses[12].

An obvious possibility, however, was that ChatGPT's creators, OpenAI, might have grown tired of stories on social media about ChatGPT's inability to answer questions about fruit boxes. And that they had now included the correct answers in the texts being trained on. Just as it also seemed likely that American law students had discussed exam questions on social media, which had then been used to train the language model on. As we all know, when you are taking an exam, it is easier to remember answers to a smaller number of known questions and then repeat these answers, than it is to understand an entire field and, on that basis, answer questions you have never heard before.

Limits to how far you can go with large language models?

That said, large language models are, of course, getting a little bit better all the time. OpenAI doesn't tell us exactly what they're working on to introduce more human-like "reasoning" into the models[13]. But it seems plausible that it could involve dividing the process of giving answers to difficult problems into smaller steps, "chains of thoughts", trying different reasoning "paths" and making different kinds of validations along the way[14]. If a user, i.e. millions of users, is satisfied with a chain of reasoning, the system can remember that, just as the system can take note if users are dissatisfied with a certain reasoning chain. And in that way improve.

It was with some excitement that many of us in the fall sat ready to test the ChatGPT model o1[15]. Fruit boxes are apparently popular when it comes to coming up with "never-before-seen problems", so in a popular problem, ChatGPT o1 is asked to "imagine a box with 4 holes where you can stick your hands in, inside the box there is a switch at each hole. Which can be off or on. You decide which holes you want to stick your hands in. In each round you can set these switches as you want, then the box is rotated around by a super intelligent robot and lands in a new way, where you don't know what you have touched before. How many rounds do you need to be sure that all switches are either off or on".

With a piece of paper and a little time, many people will be able to come up with a strategy for what they should do to solve the problem. But. Gosh. Would ChatGPT 01 be able figure it out? Were we here faced with Artificial General Intelligence, human-like intelligence here.

When I tried, I got a long explanation from ChatGPT o1, filled with a lot of intermediate calculations and finally a very precise answer, to which was added QED. Quod erat demonstrandum.

It was like getting an assignment from a very self-confident, smug student, where you must then start investigating. Is the answer really correct?

The answer was wrong. Which would not have surprised AI researcher Subbarao Kambhampati from Arizona State University, who, with a background in automated planning, has in recent years been a constant voice reminding everyone that critical thinking is as important now as ever: Large language models cannot verify their own answers. They cannot "self-verify"[16]. If you break a difficult problem down into smaller parts, all links in the chain of arguments must be correct before you can be sure that the final answer is correct. And there are no external universal mechanisms you

can call along the way to decide whether a statement is true or false[17].

Kambhampati also does not believe that the solution is simply to make the large language models even bigger, trained on even larger amounts of data. In the hope that the models will then develop "emergent behaviours" in the form of human-like "reasoning" abilities[18]. According to Kambhampati, far too many have been far too uncritical when reporting on so-called "emergent behaviours". We must remember that the internet is huge, and many times we probably just get things back that are already out there in one form or another[19]. And therefore not original thinking. In general, Kambhampati is not a big fan of "emergent behaviours". When you build an IT system, you want to know exactly what you have built. It is not in itself praiseworthy if your IT system exhibits behaviour that you cannot explain[20]. An AI system that is supposed to be responsible for railway operations must not hallucinate on the job. Certainly, people would become uneasy if physical robots around them suddenly started doing completely unexpected things.

Towards AI systems with their own models of the world.

And Kambhampati is not the only one who is somewhat reserved about how far one can actually get with large language models. According to AI researcher, and Chief AI Scientist at Facebook, Yann LeCun, we are still very far from AI systems that can reason and plan at a human-like level. According to LeCun, language models predict and understand the world (i.e. texts) in one dimension, while AI models that work with images and video understand the world in 2 dimensions. But to truly understand the world, and be able to think, one must be able to make simulations in 3 dimensions. One must have a "world model" in which one can test, simulate, actions in the world before actually doing anything[21]. With such simulations one can explain why one did a certain action (the alternatives were worse). And one can be surprised when one sees something that contradicts one's model of the world (A sign that one should be extra attentive, and ready for more unexpected things to happen).

There are of course many problems that need to be solved before we get to the point where we live in a world with robots walking around guided by their own "world models". LeCun points out, for example, the problem of dividing a “headline plan” into small sub-plans. Sure, if you're going from the US to France (LeCun is originally from France, but now lives in the US), you'll have to fly, but you'll also have to have a plan to get to the airport, and before that you'll have to get out of bed in your home in the US. Even with all of Facebook's resources at your disposal, it's not that easy to write a program that can solve these kinds of problems, in a "general" way[22].

"General intelligence" is just not easy. It's not just about having several "narrow intelligences", but also about getting these "narrow intelligences" integrated together into a whole. Human brains are not only equipped with language understanding and language generation in the brain's Wernicke and Broca areas. These brain areas have connections to many other areas, such as the prefrontal cortex, which can contribute to planning and decision-making.

Perhaps Artificial General Intelligence is not just around the corner. And many statements from marketing departments around the world will probably eventually be seen in the same light as we now see the somewhat optimistic statements made by AI researchers Simon, Shaw and Newell in 1957. When they introduced the world to GPS (General Problem Solver), a computer program that already back then, in 1957, promised to be a universal problem solver.

That said, it is worth mentioning that although Kambhampati is skeptical of many statements about large language models, he never fails to praise everything the models are good at, idea generation and much more. LeCun is of course quite adamant that there is still a long way to go before we have AI systems that can make simulations in their "world models", and thereby understand the world in a human-like way. But according to LeCun, it is still possible that we end up in a future world where we all have a large staff of personal asssitants (AI agents equipped with AGI) employed to help us with whatever we want them to help us with.

According to LeCun, people worry too much about where such a story will end. The intelligent

agents will not appear from one day to the next. Certainly, most people do not go around today being worried about overpopulation on Mars[23]? We must take the problems in relation to where we actually are. Just a little faith in progress is needed. There is still time to work on a good version for the future. We humans would like to understand more about what intelligence is. Both the machines' and our own. And most of us do believe that intelligence can be a good thing. Sure, there is much work ahead, but according to LeCun, it is not all bad.

And if he thinks so, surely the rest of us can also think so?

Simon Laub

December 2024.

Notice:

This english version of the article was ***automatically*** translated by *Google translate* from the original Danish version to English in December 2024.

1 https://www.forbes.com/sites/robtoews/2024/03/10/10-ai-predictions-for-the-year-2030/
2 https://www.reuters.com/business/autos-transportation/general-motors-drop-development-cruise-robotaxi-2024-12-10/
3 https://openai.com/12-days/
4 https://arcprize.org/arc
5 https://arstechnica.com/information-technology/2024/12/openai-announces-o3-and-o3-mini-its-next-simulated-reasoning-models/
6 https://www.wired.com/story/generative-ai-will-need-to-prove-its-usefulness/
7 https://finance.yahoo.com/news/report-claims-openai-burned-8-105046378.html
8 https://www.brightworkresearch.com/first-ai-winter-and-what-the-lighthill-report-said-about-progress/
9 https://openai.com/index/chatgpt/
10 https://www.wired.com/2014/10/future-of-artificial-intelligence/
11 https://ollama.com/library/llama2-uncensored
12 https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html
13 https://www.youtube.com/watch?v=OXwGp9YeuBg
14 https://x.com/rao2z/status/1834354533931385203
15 https://youtu.be/tMWMuJF-JFo?si=c0vTebJydL31ygxo
16 https://arxiv.org/abs/2402.08115
17 https://arxiv.org/abs/2405.04776
18 https://arxiv.org/abs/2206.07682
19 https://arxiv.org/abs/2307.02477
20 https://youtu.be/y1WnHpedi2A?si=cxPp3wKF-BopG1QQ
21 https://techcrunch.com/2024/10/16/metas-ai-chief-says-world-models-are-key-to-human-level-ai-but-it-might-be-10-years-out/
22 https://www.youtube.com/watch?v=4DsCtgtQlZU
23 https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/