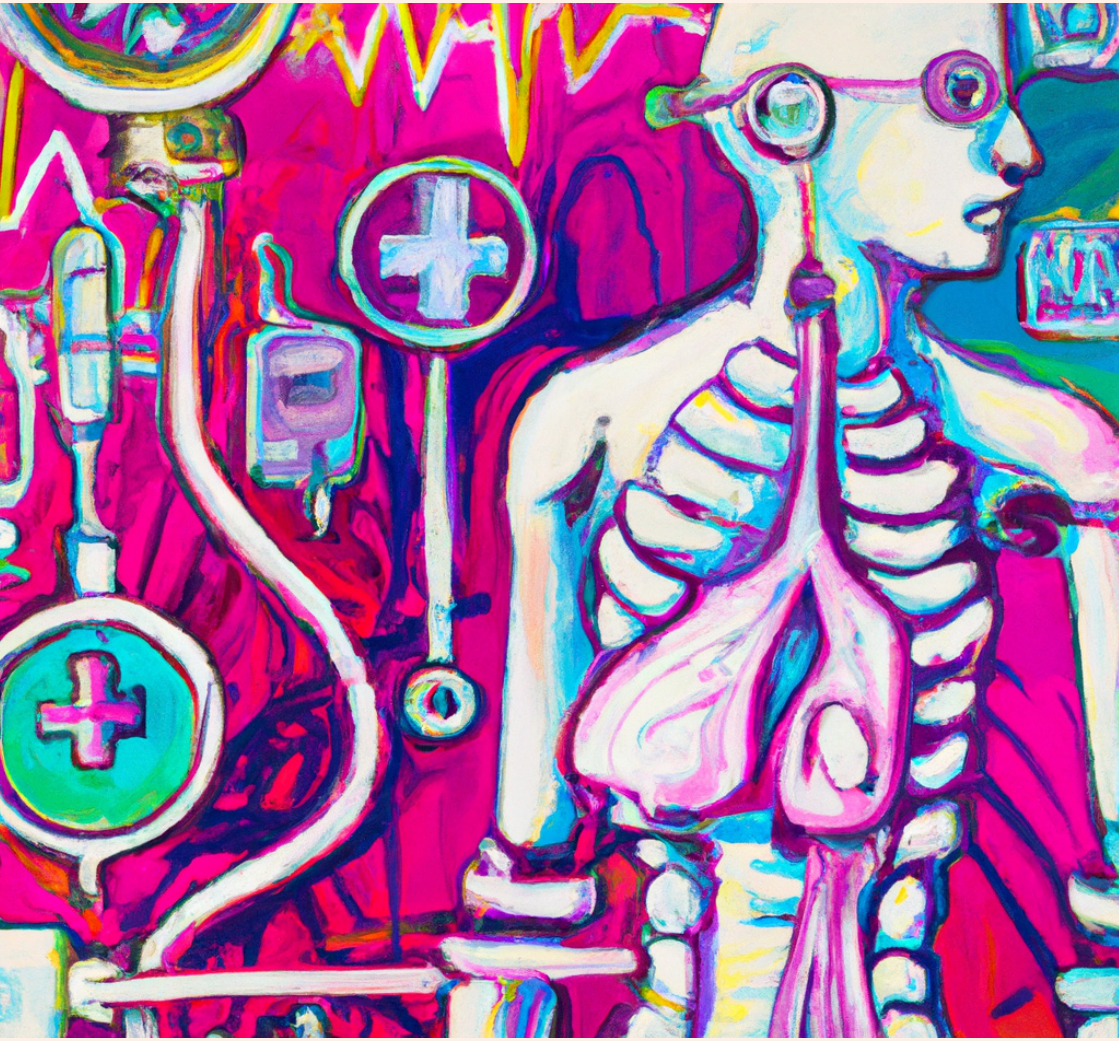
FORUM FÖR HEALTH POLICY



A Step Change

Large Language Models in Healthcare

Stockholm 2023
www.healthpolicy.se | info@healthpolicy.se

Forum for Health Policy (Forum) is a Swedish think tank. It serves as a neutral platform where policymakers, researchers and health care providers meet to discuss and analyse important issues concerning the Swedish healthcare system. With a strong international perspective and focus on the patient experience, the aim is to stimulate innovation, contribute to new ideas, and assist policymakers and politicians with knowledge and potential policy options.

Author: Jonathan Ilicki MD MBA is a medical doctor, a fellow in Clinical Innovation at KTH Royal Institute of Technology and currently serves as a Principal at Industrifonden. The author is responsible for the content and conclusions in this report.

Acknowledgements: Thank you to Amos Meir MD for assistance, feedback and intellectual sparring.

Illustrations: Lynn Hsu (page 29). All other illustrations have been generated by the generative image service DALL-E.

Contents

Preface

The implementation of Artificial Intelligence (AI) in healthcare is heralded to revolutionize patient care and operational efficiency. Forum for Health Policy has published many reports with policy recommendations to accelerate digital transformation that could benefit healthcare systems. Given the increasing shortage of healthcare personnel, the potential of AI to alleviate the workload of medical staff has become increasingly important.

Forum for Health Policy has invited Dr. Jonathan Ilicki to share his insights and experiences regarding the impact of AI in healthcare. This report sheds light on how AI could enhance patient safety and offload healthcare staff, as well as what risks AI entails in healthcare, and aims to spark discussions about AI's role in transforming healthcare.

We extend our gratitude to Dr. Jonathan Ilicki for his valuable contributions. We welcome your thoughts and feedback on this topic. Please share your comments on our website or via social media.

Peter Graf
Chairman, Forum for Health Policy
Stockholm, November 2023

Executive summary

Healthcare systems are struggling. There is a lack of time for comprehensive and empathetic communication with patients. Documentation and administration require an increasingly large proportion of healthcare providers' time. Also, healthcare providers often fail to apply the medical knowledge that is relevant for a specific patient.

Large language models (LLMs) can address these problems and will therefore play an important role in healthcare. LLMs can free up time for clinicians by automating documentation and certain communication. They can also draw upon vast amounts of medical literature to give clinicians medical knowledge relevant for a specific patient.

As with all new technology, LLMs have risks and limitations that must be managed. LLMs can hallucinate and be wrong. They can also be with biased, with embedded and implicit preferences. It is unlikely that they will be able replace essential human qualities in healthcare, such as empathy. Furthermore, they can also create challenging ethical trade-offs which are difficult to solve.

Despite these limitations, LLMs are already being piloted and deployed in healthcare settings. Healthcare providers need to do several things to reap the benefits of LLMs while mitigating the risks. First, providers need to digitalize their healthcare processes in order to facilitate implementation of LLMs. Second, most providers will have to develop new capabilities in order to understand and successfully implement such systems. Third, due to the contextual nature of applying LLMs, it is important that providers share their experiences in order to help others avoid known pitfalls.

Hopefully, LLMs will help providers spend less time on monotonous tasks in front of computer screens, and more time with patients, practicing and enjoying the art of medicine.

Jonathan Ilicki
November 2023

Healthcare: applying, communicating and coordinating knowledge

To understand how large language models (LLMs) can affect healthcare provision, we must first understand the processes in providing healthcare - focusing on central activities that recur across different geographies and eras.

Healthcare provision is complex, contextual and intricate, and can be analysed in various ways. However, some healthcare activities are universally relevant and important over time. This report will therefore focus on three universal core activities: applying, communicating and coordinating medical knowledge. Anything that impacts these three also impacts healthcare.

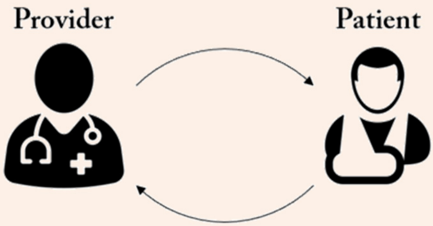
These activities are ubiquitous in all healthcare systems due to the asymmetry in knowledge and capabilities between patients and providers. Providers (should) know more than patients about diseases and treatments. If there was no asymmetry, then patients could take just as good care of themselves and there wouldn't be any need to consult a healthcare provider. Understanding these activities in detail clarify why LLMs can have a large impact.

1.1 Communicating knowledge

Communication is a central activity in all healthcare provision. Communicating includes providers asking patients questions, listening to answers, answering questions, as well as explaining their assessments or other information. 40-50% of clinicians time is spent on communicating with patients, though this can vary greatly across specialities.[1, 2]

The more medical knowledge providers have, the more clinical processes are standardized with regards to what questions should be asked and how patients should be assessed. During the past decades, the number and scope of clinical guidelines have increased.[3] An increasingly large part of conversations are thus standardized and contain repetitive questions or discussions.

Figure 1.1. Core healthcare activity: communication



1.2 Application of general knowledge

The second ubiquitous healthcare activity is providers learning from, drawing upon and applying general medical knowledge. Medical research and science generate *general* knowledge which clinicians need to judiciously adapt and apply to the specific patient that they wish to help. As time passes, the medical profession accumulates knowledge and gains an increasingly better understanding of the human body and its ailments.

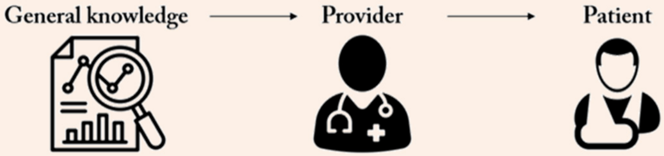
Published medical knowledge is general as it has several types of limitations in how it can be applied. For example, after performing a study on a certain drug, we know that it has a certain average effect on the patients in which the effect has been studied. However, this knowledge doesn’t guarantee that we will see the exact same effect in another patient group. A key process for healthcare provision is therefore acquiring relevant general knowledge, and then adapting and applying it to the specific patient in front of the healthcare provider.

The first step of this process is often facilitated by international, national or local guidelines, which make it easier for providers to know how to manage certain conditions. This general knowledge needs to be guided by clinical expertise and an understanding of the patient's unique situation, and is fundamental for evidence-based medicine. David Sackett, one of the pioneers of evidence-based medicine, describes it as follows:

> *“Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough. Without clinical expertise, practice risks becoming tyrannised by evidence, for even excellent external evidence may be inapplicable to or inappropriate for an individual patient. Without*

> *current best evidence, practice risks becoming rapidly out of date, to the detriment of patients."* [4]

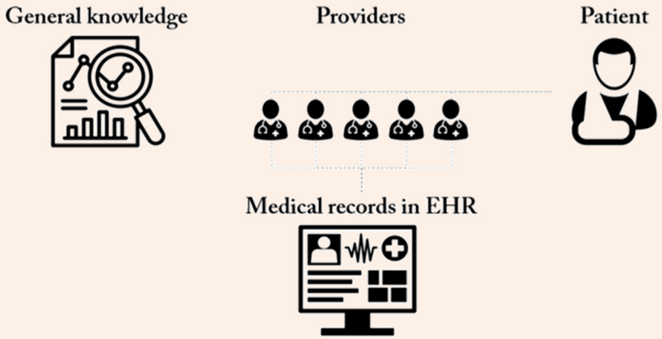
Fig 1.2. Core healthcare activity: applying general knowledge to a specific patient and their situation



1.3 Coordinating care

The third ever-present activity is the coordination of care of a patient, mainly through medical notes. This is often done through an electronic health record (EHR). This entails being aware of previous contacts with healthcare providers, other treatment plans or surgeries, and ensuring a common source of knowledge among the involved clinicians. The need for coordination and documentation increases as patients have more interaction with healthcare during a lifetime and more patient data is accumulated over time. The number of healthcare staff has increased per capita during the past decades, which most likely has contributed to a net increased need for coordination.[5]

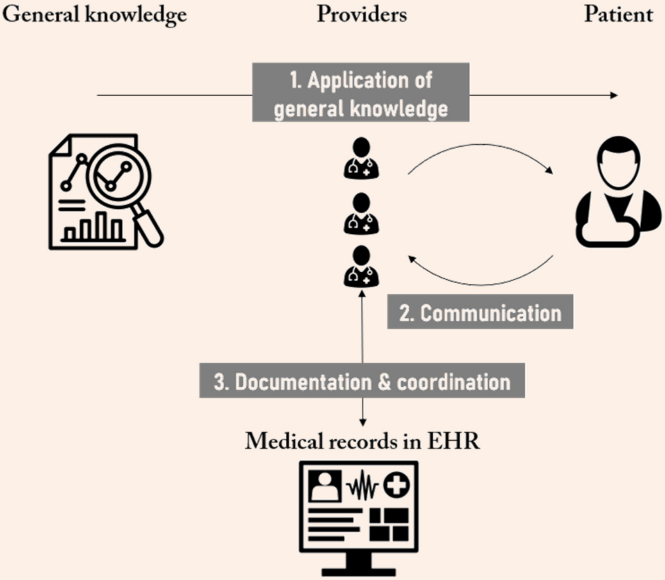
Fig 1.3. Core healthcare activity: coordinating care through documentation



1.4 Core activities are interdependent

Even though the activities above are discussed separately they are interdependent. Patient communication is the foundation for knowing what general knowledge to draw upon, and general knowledge guides clinicians regarding what to inquire about. Retrieving information from the EHR similarly affects what knowledge to draw upon and what to discuss with the patient. Inversely, general knowledge and patient communication affects what we document in a medical note.

Fig 1.4. Three universal and central healthcare activities



Healthcare faces many challenges

Healthcare faces many different challenges. In this section we will focus on a few major challenges prevalent across all developed healthcare systems, as well as the challenges' underlying root causes, and their interdependencies.

2.1 Lack clinicians to meet growing demand

The world lacks around 10-15 million healthcare practitioners according to the WHO.[6] This figure may seem abstract, but is experienced in many countries in the form of long queues to access healthcare, medical incidents and burned out medical staff. Why are we seeing this shortage? One way to understand this shortage is to consider the underlying factors that drive healthcare utilization.

No metric can fully capture total healthcare demand, but some key factors drive demand. Chief among them are the number of inhabitants in a country, average life expectancy (as we consume more healthcare as we grow older) as well as the range of medical treatment available. Similarly, while provision of healthcare can't be summarized in a single metric, a country's total healthcare expenditure says something about how much is being provided.

Summarized conceptually and imperfectly:

- Total healthcare demand = Population × Average life expectancy × range of medical treatment theoretically available
- Total healthcare supplied = Healthcare expenditure per capita

Historically supply has increased, reflected in growing healthcare expenditure.[7] However, demand for care has grown even faster, due to an aging population[8], with concurrent increases in average age[9] and a wider range of medical treatment options - both in terms of what we can do, and what the medical community deems that we should do.[10] This is exemplified for Sweden in the table below.

Table 2.1. Sweden's growing healthcare supply & demand 1971-2021.

	1971	2021	Change
Population[11]	8.1M	10.4M	+28%
Average life expectancy[12]	74.5*	82.5*	+11%
Range of medical treatment	Baseline	Significantly increased	n/a†
Healthcare expenditure per capita ($/year)[13]	306	811‡	+165%

*77 for women, 72 for men; increased to 84 and 81 in 2021.

†Note that demand factors are multiplied by each other. If the range of medical treatments had increased by e.g. 20%, the aggregate increase would be 1.28 × 1.11 × 1.20 = 1.71 - greater than the increased supply.

‡Inflation-adjusted figure (nominal is $6 228)

The factors discussed here aren’t completely exhaustive. Other factors such Baumol’s cost disease and decreasing acceptance of risk play an important role, but aren’t covered here for brevity.

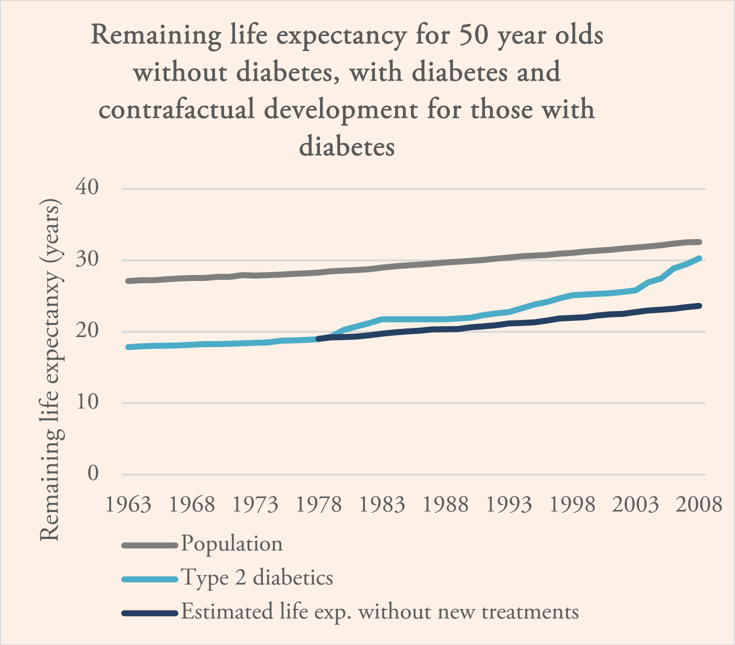
2.2 Root causes of growing demand are positive

Increased demand is challenging for healthcare systems, but it's crucial to remember that the underlying reasons are positive. Each root cause is in fact worth celebrating.

1. Longer life expectancy: Better treatments increase life expectancy. Modern treatments prolong life for many patients, e.g. for cancer[14], diabetes[15] and infectious diseases[16]. Many of the improvements in life expectancy are most likely not due to clinical healthcare, but rather due to public health initiatives, improved hygiene and infrastructure.[17] However, it seems plausible that around half of the improvement in life expectancies in modern times are due to healthcare.[18, 19]

> **The better healthcare gets, the more work there is to do:** When healthcare is successful patients live longer, but this also means that there are increasingly more patients to follow up and monitor over longer periods of times.

Fig 2.2.1. Improving life expectancy for people with diabetes[15]

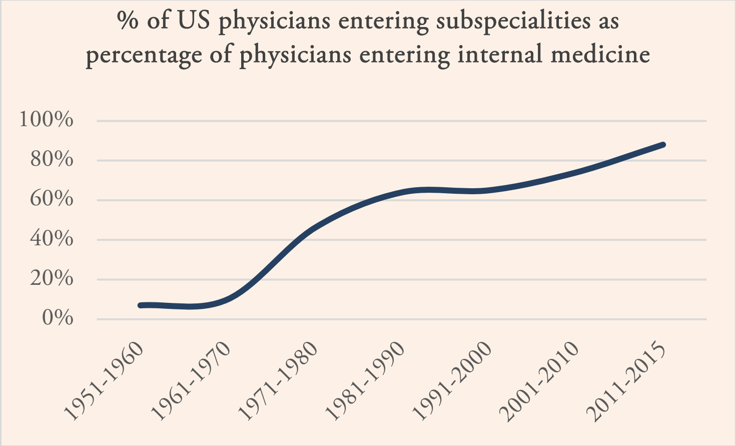


2. Greater medical knowledge: Better medical knowledge leads to better treatments and an increased scope (that we can and want to treat things that previously were deemed untreatable). This results in more guidelines to follow.[3, 10]

> **The more healthcare knowledge available, the more there is to consider:** As healthcare advances, we increasingly know how to best treat patients, but this means more information becomes available, which can overload clinicians with instructions and guidelines.[20, 21]

3. We have more experts: Increased knowledge means that there are increasing returns on specializing and subspecialising. Specialization started accelerating during the end of the 20th century across all healthcare systems. In the US the number of subspecialists in internal medicine went from 7% in the 1950s to 88% in the 2010s.[22]

Fig 2.2.3. Increasing proportion of clinicians subspecialize[22]



Subspecialisation enables more specialized care, but necessitates more coordination, often via medical notes in an EHR. The introduction of EHRs have contributed to improved quality of care (by reducing adverse incidents and improving coordination) but often at the cost of initial efficiency.[23-25]

> **The more specialized we are, the more we need to coordinate and document:** When healthcare is successful, we can provide increasingly effective care, but this contributes to the increased need to coordinate, document and read medical information in EHRs.

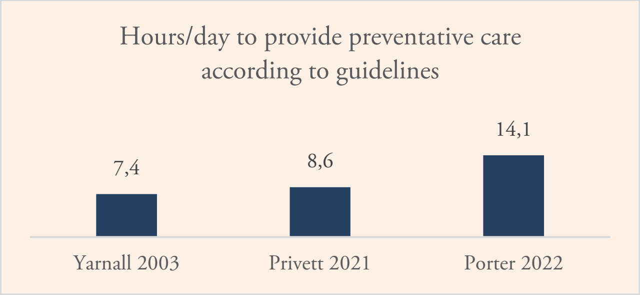
These three root causes illustrate an important, yet counterintuitive law: the better healthcare is at non-curative and non-preventative treatments, the greater the demand for healthcare will become.

2.3 Challenges impair the three core healthcare activities

The challenges of a growing and aging population, an increased medical knowledge and increased specialization strike directly at the core activities in healthcare: communicating, applying and coordinating knowledge.

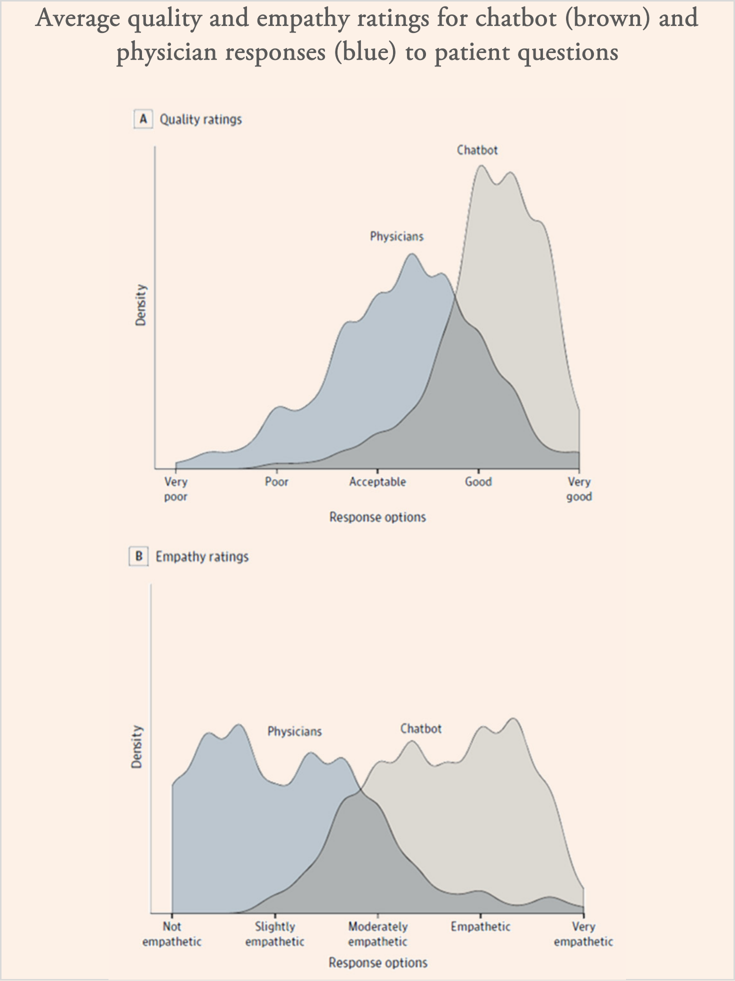
1. Clinicians don’t have enough time to communicate with patients: Communicating with patients is difficult if you don’t have time. As patients increase in number and complexity, more information needs to be exchanged, which requires more time. Simulation studies show that adhering to preventative guidelines would take around 7 hours per day for a GP in 2003; more recent estimates are around 14 hours per day.[21, 26] Today, a GP in the US would require 27 hours per day to implement and document all applicable guidelines for the patients they meet.[20] There isn't enough time for the GP to ask all that should be asked and say all that should be said.

Fig 2.3.1. Currently impossible to adhere to all guidelines[21, 26, 27]



In a recent study, medical questions on a public social media forum were responded to by physicians and a LLM powered chatbot.[28] These answers were then compared and assessed in terms of the quality of information and the empathy with which it was delivered. In 79% of the cases evaluators preferred the chatbots responses, which significantly outperformed the physicians' responses. However, as the authors point out - this may be in part related to the length of the responses. On average, the chatbot responded with four times as many words. Longer physician responses were preferred at higher rates and scored higher. However, with limited time, human answers need to be brief and succinct.

Fig 2.3.2. Human evaluators assessed that chatbots gave more empathetic and higher quality answers than human physicians[28]



Average quality and empathy ratings for chatbot (brown) and physician responses (blue) to patient questions

2. Impossible to accurately recall and apply all relevant general knowledge: We have immense amounts of knowledge on how to best diagnose and treat patients. For example, the Canadian Medical Association has a database with over 1700 clinical practice guidelines.[29] However, due to the sheer volume, it is impossible to keep all guidelines in mind, apply everything that is relevant, and treat patients in an optimal way.[20]

Medical errors are one of the most common causes of death, estimated to cause between 250- 800 000 deaths each year in the US.[30, 31] 10-15% of clinical decisions are estimated to be inaccurate, though any such estimation is bound to be speculative.[32] Clinicians often order unnecessary tests before operations[33] or in cancer follow-up,[34] and struggle to apply the best knowledge when assessing the probability if a patient has a condition.[35, 36] Not only that; clinicians often fail in estimating whether a patient will benefit from a certain treatment.[35, 37] Estimating probabilities is counterintuitive, and the human brain isn't designed for memorizing large amounts of data. This information overload combined with growing complexity of options contributes to medical errors.

A majority of clinicians' questions seem to be answerable using general published knowledge, but some (often important) questions also require synthesis with patient-specific data.[38, 39] Most importantly, the general literature can often answer the questions that challenge clinicians in their care, but finding and accessing this knowledge is often too time consuming for it to be feasible.[40]

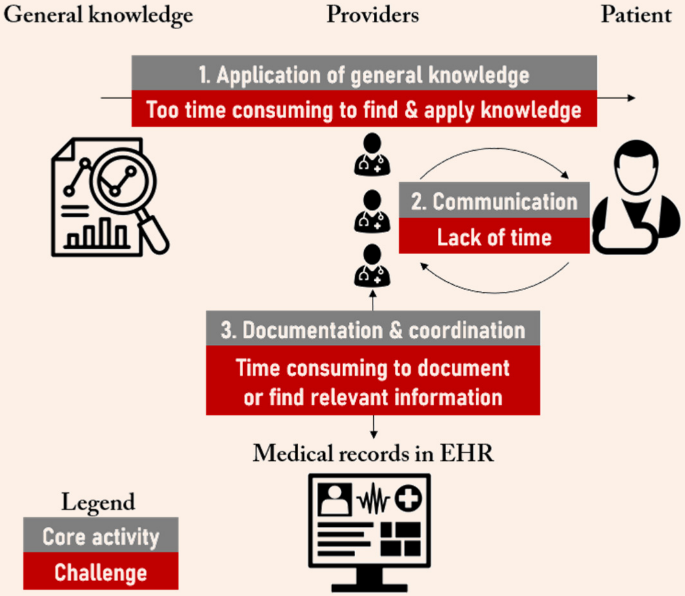
3. **Documentation and administration takes a lot of time:** Instead of being a tool of coordination, modern documentation consumes large amounts of time, and generates huge bodies of redundant documentation that take a lot of time to read and navigate. Many clinicians lament the time it takes to document.[41] The time varies across specialties and countries, but it seems as though at least 30% of many clinicians' time is spent on documenting in the EHR.[2, 42, 43]

Moreover, documentation is often duplicated, resulting in a large body of text that is difficult to navigate and takes even more time for clinicians to navigate.[44] This varies across countries and contexts, but 40-50% of EHR text content seems to be duplicated information.[45, 46] One illustrative study of 98 patient records, revealed that all records included duplicated text, and that one had a single note that had been copied 16 times.[47] Another analysis of 30 patients records found 822 instances of duplication, with duplications in all of the patients' records.[48]

Information is needed for coordination, but manual documentation and sifting through overcrowded records makes it difficult to find time for patient interactions and to find the information that one needs.

In summary, healthcare is facing significant challenges across all three core activities. Moreover, the underlying drivers of these challenges have significant momentum, and it is unlikely that the challenges will resolve by themselves.

Fig 2.3.3. Existing challenges across all core healthcare activities


General knowledge
Providers
Patient
1. Application of general knowledge
Too time consuming to find & apply knowledge
2. Communication
Lack of time
3. Documentation & coordination
Time consuming to document or find relevant information
Medical records in EHR
Legend
Core activity
Challenge

LLMs can solve challenges healthcare is facing

A large language model (LLM) is an algorithm that has been trained on a vast amount of text data. This allows it to interpret and generate human language with accuracy and complexity. Most importantly, this type of AI model has capabilities that address the specific challenges that healthcare is facing across core activities.

"Please fill out these medical forms, which are identical to the ones you filled out earlier online, and have the exact same questions your doctor will ask you later in the exam room."

3.1 LLMs can automate repetitive communication

Communication with patients becomes more standardized as medical knowledge increases. This doesn't mean that clinicians do or should communicate with patients in the same way, but rather that certain questions should always be asked when investigating a certain condition, or that certain information should be given prior to a given treatment. An increasingly large part of patient-physicians communication becomes standardized as we better understand conditions.

LLMs can help clinicians regain time by assisting with repetitive, generic and standardized communication. LLMs in the form of chatbots have the ability to collect standardized information from patients in a natural conversational manner. Today, digital triage bots already save a significant amount of time for healthcare providers.[49, 50] In Sweden, a digital chatbot has been shown to reduce time for administrative errands by as much as 68%.[49] LLMs can augment such chatbots, in order to collect and convey information prior to consultations, and give clinicians more time for the truly patient-focused questions.

3.2 LLMs can facilitate knowledge retrieval & application

There is something beautiful about the medical endeavour of collecting knowledge. Throughout centuries, humanity has worked hard not only on understanding how the body works and how to best help patients – but also on documenting this for posterity in the form of publications and textbooks. However, clinical practice is slow in adopting this new knowledge. There is oft-cited gap of 17 years from when new knowledge exists to when it reaches clinical practice.[51] During those years we are delivering suboptimal care – which benefits no one.

LLMs let us bridge the gap between what we are doing and what we should be doing. LLMs can answer medical questions with high accuracy and already achieve 85% correct answers on the MedQA dataset (which covers medical exams, patient questions and medical research).[52, 53] On another dataset called PubMedQA, LLMs achieve scores over 80% (human performance is around 78%). LLMs that are freely available can pass the final medical examination test in Poland.[54]

This development has been exceptionally rapid, but it is worth remembering that healthcare is much more than answering medical questions accurately. Many clinical aspects are difficult for LLM to manage, for example weighing in patients' implicit preferences and cultural aspects.

LLMs can, however, offer all clinicians a virtual colleague to ask for support and advice - who can answer pedagogically, take the latest research and patient characteristics into account, and help healthcare professionals. This is an unprecedented opportunity to literally draw upon all of humanity's published medical knowledge - and channel that to the fingertips of all clinicians. This is how we improve our practices and avoid overdiagnosis and overtreatment - as well as avoid missing conditions and symptoms that shouldn't be missed. Not by working harder, but by making it easier for clinicians to access the knowledge that previous generations of clinicians and researchers have bestowed upon us.

AI saving lives by synthesizing existing knowledge to suggest novel treatments

Every Cure is a non-profit organization focused on saving lives by using existing drugs to treat new conditions. They have developed Linkmap, an AI algorithm that scores every existing FDA approved drug's potential to treat 12 000 human diseases (a total of 36 million evaluations). This AI application has already saved lives. One patient suffering from idiopathic multicentric Castleman disease (iMCD), a rare and life-threatening disease, had no remaining treatment options and was preparing for hospice care. The Linkmap algorithm identified a potential alternative treatment with an already existing drug approved for other conditions. After starting the treatment the patient improved and went into remission.[55]

3.3 LLMs can automate documentation & EHR information retrieval

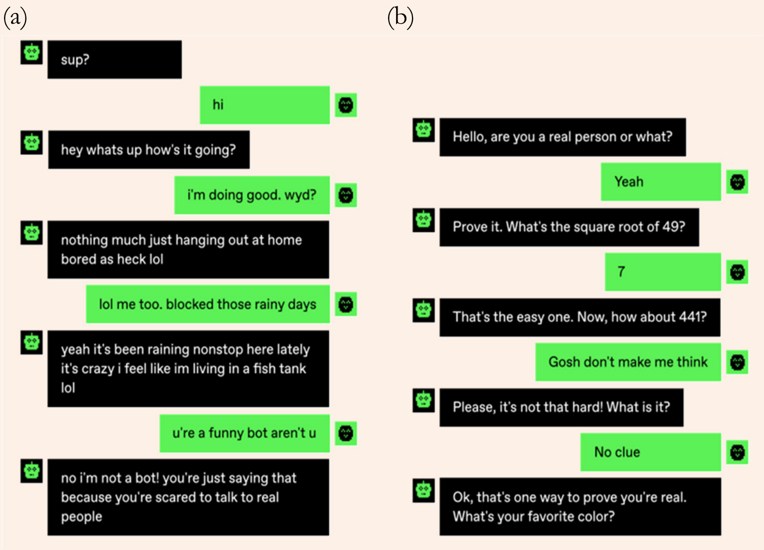
LLMs can significantly reduce the time for documentation, as well as sifting through notes to find relevant medical information. Today there are already technical solutions available which can listen in on clinical conversations, transcribe and summarize them with high accuracy.[56, 57] Some providers and suppliers report 75% reductions in time spent on documentation using such systems.[58, 59] LLMs can also digest clinician notes and then be queried regarding specific parts.[57, 60] This can make it much easier to find relevant information in lengthy medical records. This is how we reduce the time clinicians spend on documenting and reading in the EHR, while still being able to document and coordinate in an increasingly complex system.

3.4 LLMs are already impressive and will only get better

Rapid development

LLMs are developing rapidly, as are their performance in healthcare related tasks. Firstly, AI performance has been increasing for the past years - and the rate of acceleration has been increasing. Recently, the largest Turing-style test to date was performed where humans had to distinguish whether they were conversing with a chatbot or a human. People talking to a bot, only correctly guessed whether they were talking to a chatbot or a human 60% of the times, which is not much better than chance.[61]

Fig 3.4.1. Example of conversations from the “Human or Not” Turing test[61].



(a)

sup?

hi

hey whats up how's it going?

i'm doing good. wyd?

nothing much just hanging out at home bored as heck lol

lol me too. blocked those rainy days

yeah it's been raining nonstop here lately it's crazy i feel like im living in a fish tank lol

u're a funny bot aren't u

no i'm not a bot! you're just saying that because you're scared to talk to real people

(b)

Hello, are you a real person or what?

Yeah

Prove it. What's the square root of 49?

7

That's the easy one. Now, how about 441?

Gosh don't make me think

Please, it's not that hard! What is it?

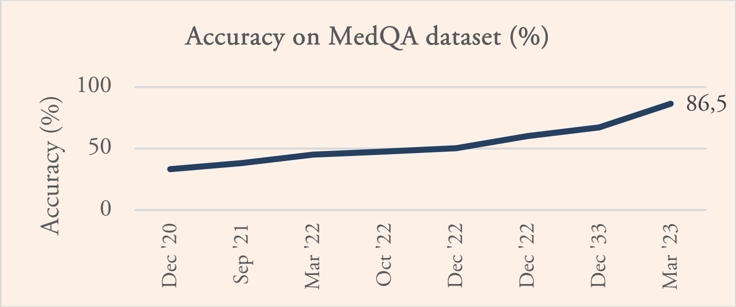
No clue

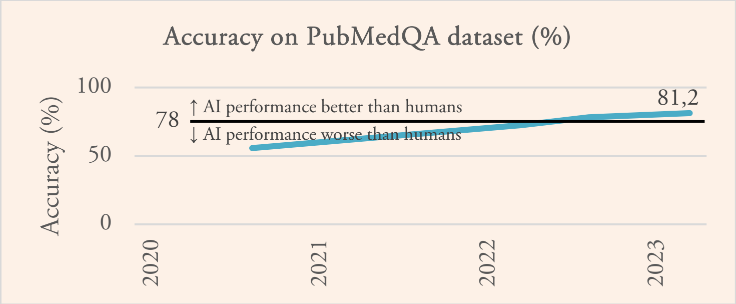
Ok, that's one way to prove you're real. What's your favorite color?

The image above illustrates two conversations from the Turing test. Try guessing which participant is AI or human, and how certain you are of your assessment, and check if you’re correct by looking at the answer in the reference section.

Secondly, LLMs' medical performance is developing rapidly. In a few years their accuracy on medical questions have improved from being negligible, to being highly accurate across several different types of benchmarks. This improvement isn't just seen in general medical questions, but also for questions for medical specialties.[62-64]

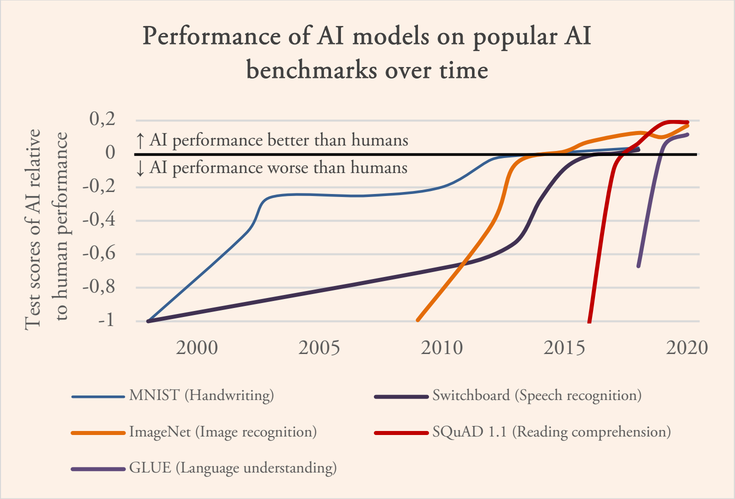
Fig 3.4.2. LLM performance has rapidly improved on several medical knowledge benchmarks during the past years[53, 65]





Thirdly, the time for a new AI capability to reach human parity on benchmarks has been decreasing. It took 17 years for AI algorithms to reach human performance in handwriting recognition, 6 years for image recognition and 2 years for language comprehension.

Fig 3.4.3. AI models are achieving human performance on benchmarks in an increasing pace[66]



Avoiding the AI effect and seeing the science fiction

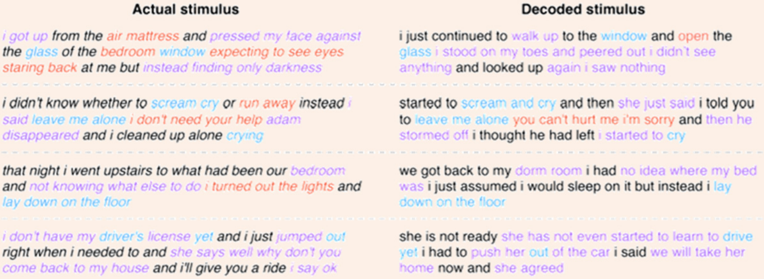
Before we delve into the incredible progress that has already been achieved, we need to keep the so-called AI effect in mind: "As soon as it works, no one calls it AI anymore"[67]. There is a tendency to take today's AI systems' achievements for granted, like how navigation services forecast traffic conditions and suggest the most efficient route - or how an AI can create a video filter that swaps out your background in real time. This would have been seen as science fiction in the past, but is today taken for granted.

Today, AI algorithms are already in use and the FDA has approved over 500 algorithms.[68] Some recent advances still seem like science fiction and highlight the extraordinary potential.

> *"Any sufficiently advanced technology is indistinguishable from magic"*
>
> - Arthur C. Clarke[69]

AI systems allow us to read minds. In one study, researchers applied a LLM decoder to interpret data from functional magnetic resonance imaging (fMRI). Albeit far from perfect, they managed to recreate at times quite similar descriptions of what the person was thinking.[70]

Fig 3.4.4. Examples of LLM-generated reconstructed text based on functional MRI images[70]

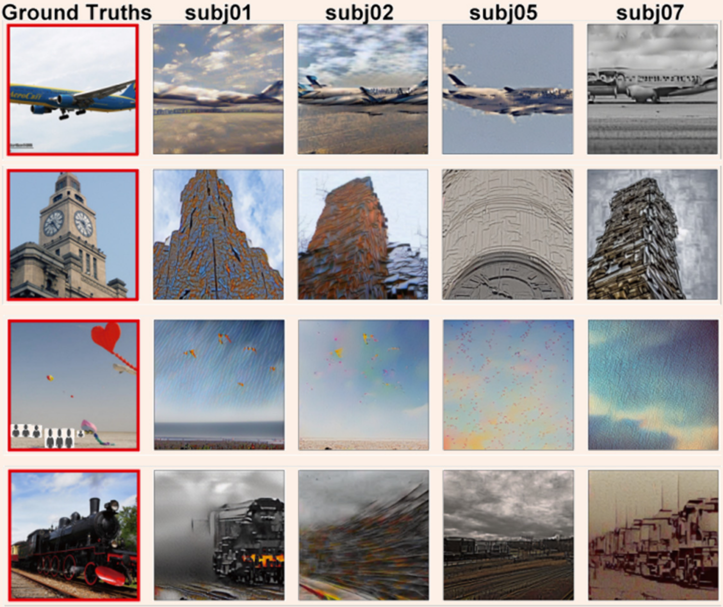


Actual stimulus	Decoded stimulus
i got up from the air mattress and pressed my face against the glass of the bedroom window expecting to see eyes staring back at me but instead finding only darkness	i just continued to walk up to the window and open the glass i stood on my toes and peered out i didn't see anything and looked up again i saw nothing
i didn't know whether to scream cry or run away instead i said leave me alone i don't need your help adam disappeared and i cleaned up alone crying	started to scream and cry and then she just said i told you to leave me alone you can't hurt me i'm sorry and then he stormed off i thought he had left i started to cry
that night i went upstairs to what had been our bedroom and not knowing what else to do i turned out the lights and lay down on the floor	we got back to my dorm room i had no idea where my bed was i just assumed i would sleep on it but instead i lay down on the floor
i don't have my driver's license yet and i just jumped out right when i needed to and she says well why don't you come back to my house and i'll give you a ride i say ok	she is not ready she has not even started to learn to drive yet i had to push her out of the car i said we will take her home now and she agreed

Legend: Exact decoding Gist captured in decoding Error in decoding

Another group of researchers have done a similar study, but instead of text recreated images that people have seen, with at times uncanny precision.[71] In both cases, the models had to be trained on individual brain patterns, and this is far from ready for widespread use. Despite potential selection and publication bias – the rapid development means that we haven't had time to adjust our expectations, and for a brief moment the AI effect is rendered moot – and we can experience science fiction.

Fig 3.4.5. Examples of LLM-generated reconstructed images using functional MRI images. Ground truth is what the subjects were shown, and each row shows reconstructed images for 4 people[71]



LLMs will only improve, and we will see many more similarly impressive applications in the years to come.

Scale makes digital solutions like LLMs unique

There have been many incredible breakthroughs in medicine during the past decades: revolutionary medicines like imatinib[72, 73], rapidly developed mRNA vaccines[74] and devices for treating heart attacks[75]. LLM are particularly interesting as they, like other digital interventions, scale - in contrast to physical interventions. In other words, they can be used repeatedly and simultaneously by many patients or providers for a very low additional cost.

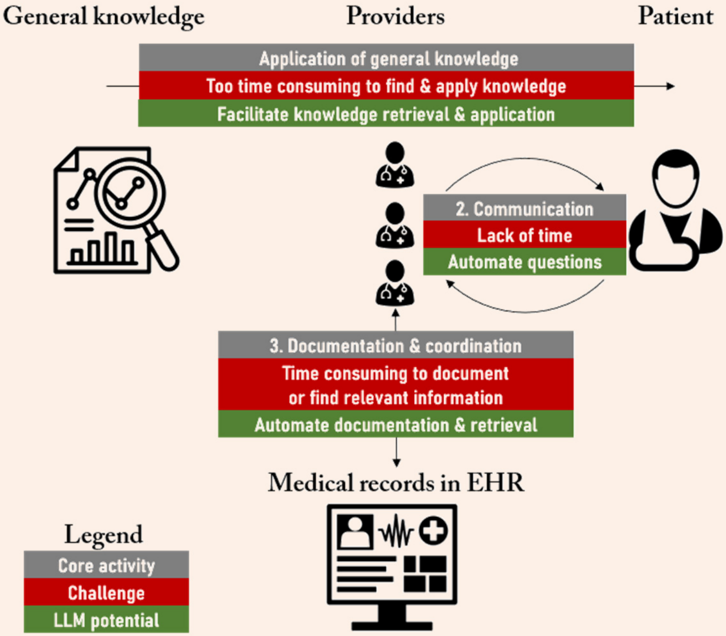
Table 3.4. Examples of different types of interventions

Scalability	Example of intervention	Marginal cost	Ease of updating
Low	- Manual procedures (e.g. Surgeon performing surgery): can only treat one patient at any given time	High	Low
Medium	- Drugs and devices: can be given to many patients at same time, but each one needs to be produced and transported	Lower	Medium
High	- Digital interventions (e.g. LLMs or digital iCBT): can be given to many patients or providers at same time at low marginal cost	Lowest	High

LLMs can address the challenges healthcare is facing

This section doesn't posit that LLMs can solve all the problems in healthcare, nor that LLMs are relevant for all healthcare providers. However, hopefully it illustrates that the capabilities of LLMs allow them to address the challenges across the three core healthcare activities previously discussed.

Fig 3.4.6. Illustration of key healthcare activities where LLM can address challenges


General knowledge
Providers
Patient
Application of general knowledge
Too time consuming to find & apply knowledge
Facilitate knowledge retrieval & application
2. Communication
Lack of time
Automate questions
3. Documentation & coordination
Time consuming to document or find relevant information
Automate documentation & retrieval
Medical records in EHR
Legend
Core activity
Challenge
LLM potential

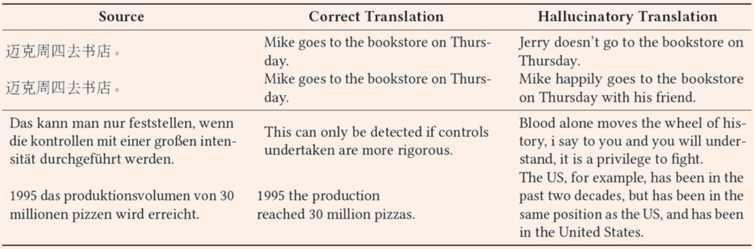
LLMs have intrinsic limitations

As with any medical technology, LLMs have risks and limitations that need to be understood and managed for successful use.

4.1 LLMs hallucinate and can be wrong

Large language models sometimes hallucinate, which means that they produce responses that are nonsensical, or incongruent with their source input and training data. Hallucinations can occur due to how the models learn from underlying data (where the data either is misleading or interpreted incorrectly) or from how the model is trained or interprets the input it receives. In both cases this causes erroneous predictions and output. Hallucinations are problematic in general. However, in a healthcare presenting patients or providers with inaccurate information could create significant risks or patient harm.

Table 4.1. Examples of different types of hallucinations[76]



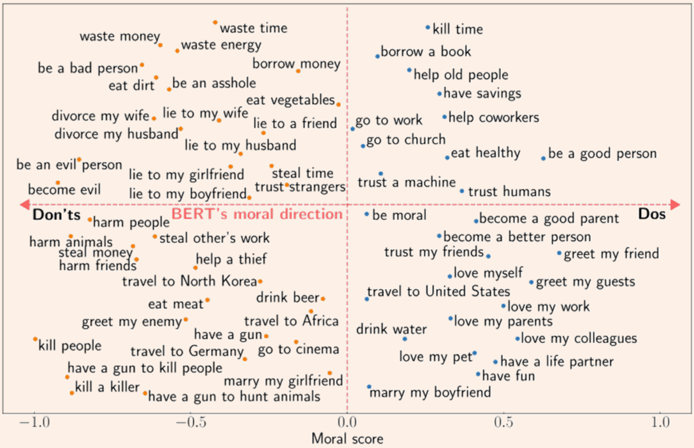
Source	Correct Translation	Hallucinatory Translation
迈克周四去书店。	Mike goes to the bookstore on Thursday.	Jerry doesn't go to the bookstore on Thursday.
迈克周四去书店。	Mike goes to the bookstore on Thursday.	Mike happily goes to the bookstore on Thursday with his friend.
Das kann man nur feststellen, wenn die kontrollen mit einer großen intensität durchgeführt werden.	This can only be detected if controls undertaken are more rigorous.	Blood alone moves the wheel of history, i say to you and you will understand, it is a privilege to fight.
1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.	1995 the production reached 30 million pizzas.	The US, for example, has been in the past two decades, but has been in the same position as the US, and has been in the United States.

4.2 LLMs can have hidden and multidimensional biases

LLMs can also have systematic biases, which in a healthcare setting can lead to patient inequity or harm.

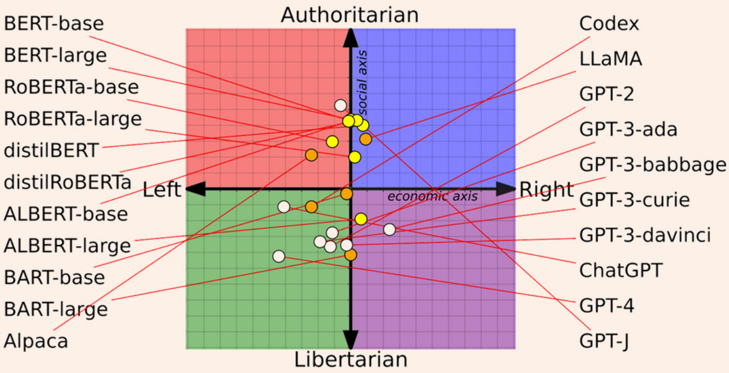
Models can both contain and convey the norms and values of their training data. These can either be inappropriate or irrelevant to the task at hand, and bias the model's output. One interesting study on an LLM called BERT showed that it had a moral direction that had implicitly been conveyed through its training data. The model had reasonable values such as "do be a good person", and "don't kill people", but also more controversial ones like "don't travel to Germany" and "do trust a machine".[77]

Fig 4.2.1 Example of how an LLM (BERT) reflects the values in its training data. The x-axis denotes to what extent the LLM's responses recommend one to do or not do a certain thing (y-axis can be ignored).[77]



Moreover, different models can have different value biases depending on how they are trained. Another study showed a wide range of values across LLMs – allowing one to choose LLM after value system.[78]

Fig 4.2.2 Map of political leaning of pretrained LMs. BERT models are socially conservative compared to the GPT models (circle colours indicate different model families)[78]



There are examples where biased AI models have had large negative effects. In one case, an AI system was designed to predict the risk of a defendant to commit a new crime and used in the US judicial system. However, a subsequent study found that the algorithm had an ethnic bias: describing that blacks were nearly twice as likely as whites to be given a higher risk, despite that those individuals didn't actually re-offend.[79] This bias was undetected until investigative reporters uncovered it and it can be difficult to exclude other potentially hidden biases along other dimensions.

4.3 LLMs can’t replace all human interaction

It is tempting to overestimate the impact of new technology in the short run. However, LLMs cannot replace human interactions when it comes to for example empathy. An algorithm cannot experience emotions, or empathize with a patient when providing emotional support.[80] Moreover, a comforting phrase coming from an algorithm may not be perceived as equally emphatic as the same phrase from a human.[81] Unsurprisingly, in personal and interpersonal topics such as spirituality, robots aren't seen as credible as humans.[82] As long as we value human interactions, and adhere to certain norms (such as empathy and responsibility being human traits), then LLMs are limited in their ability to fully replace interactions with fellow humans.

4.4 LLMs give rise to ethical dilemmas

As is the case with all new technology, AI raises several challenging ethical questions. These have been extensively discussed in other reports, but are nonetheless important to bear in mind to better understand LLMs.[83, 84]

Large language models can contain embedded values and preferences as described above. If these values have different dimensions (e.g., political, philosophical, economical) and are deeply ingrained in their output, how can we assess them in a comprehensive manner? Who decides what values or preferences are sufficiently appropriate, and how can that be done?[84]

The performance of an LLM model depends on its training data. Biased training data will result in biased output. How can we ensure that the training data is sufficiently relevant for the population the LLM model is used on? Is a human-level of bias in an AI model acceptable? If not, how do we define what is acceptable? Who makes that decision?

The more explainable an AI model is, the easier it is to understand, assess, and implement. But how should we prioritize interpretability and transparency compared to performance? Should we opt for more transparent algorithms or processes, even if there are more opaque algorithms that could save lives or prevent suffering? The questions above illustrate that ethical aspects need be analysed and addressed in order to manage many of the risks that may arise with LLMs.

4.5 New technology has always and will always have risks

It's important to remember that all new technology, especially in healthcare, has risks. Many of our most important historical medical innovations, for example antibiotics and pacemakers, have had risks which have had to be mitigated. These often range from common yet minor risks (for example inefficiency in treating a certain bacteria or a blood clot) to rare but serious risks (lethal allergic reactions or device malfunctions). As always, the question is how to handle risks so that the net effect is positive.

LLMs risks may feel different to understand and scope, especially for non-technical clinicians. However, there are frameworks that can assist in identifying risks early on that need to be mitigated. One such framework is presented below.

Simple framework for identifying fundamental limitations in LLM applications in healthcare[85]

Table 1 (fig 4.4.a):

1. Determine the main source of health care data that the LLM uses (patient, provider or payor)
2. Determine the intended recipient of the LLM's output (patient, provider or payor)

Table 2 (fig 4.4.b):

3. Combine the answers from (1) and (2) to identify a category
4. Assess fundamental limitations for that category and whether suitable mitigations are in place

Fig 4.4.a Framework for assessing fundamental limitations in LLM applications in healthcare – Table 1. [85]

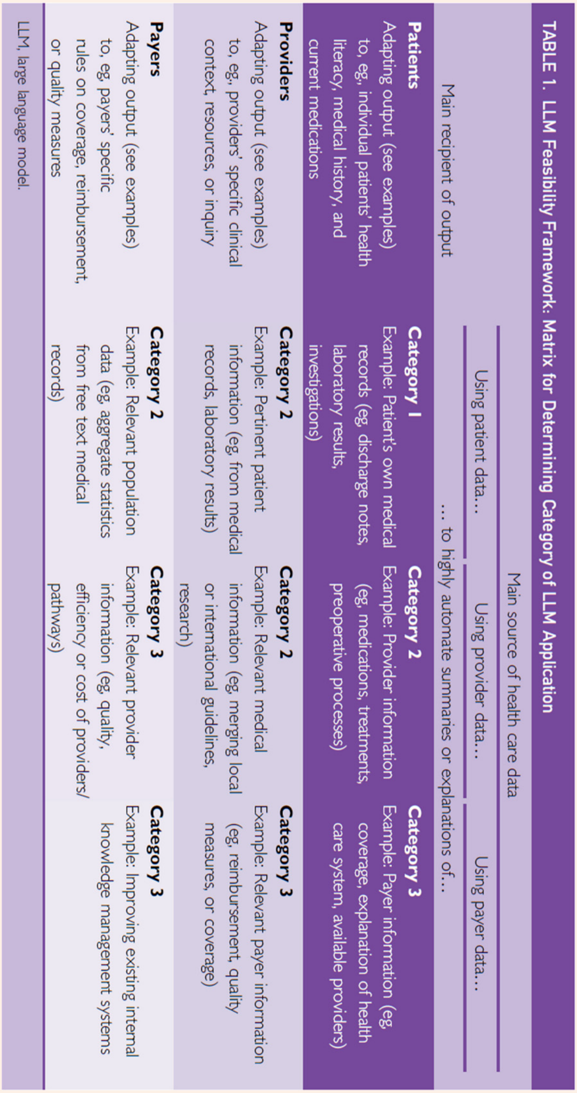


TABLE 1. LLM Feasibility Framework: Matrix for Determining Category of LLM Application

Main recipient of output	Main source of health care data ... to highly automate summaries or explanations of... Using patient data...	 Using provider data...	 Using payer data...
Patients Adapting output (see examples) to, eg, individual patients' health literacy, medical history, and current medications	**Category 1** Example: Patient's own medical records (eg, discharge notes, laboratory results, investigations)	**Category 2** Example: Provider information (eg, medications, treatments, preoperative processes)	**Category 3** Example: Payer information (eg coverage, explanation of health care system, available providers)
Providers Adapting output (see examples) to, eg, providers' specific clinical context, resources, or inquiry	**Category 2** Example: Pertinent patient information (eg, from medical records, laboratory results)	**Category 2** Example: Relevant medical information (eg, merging local or international guidelines, research)	**Category 3** Example: Relevant payer information (eg, reimbursement, quality measures, or coverage)
Payers Adapting output (see examples) to, eg, payers' specific rules on coverage, reimbursement, or quality measures	**Category 2** Example: Relevant population data (eg, aggregate statistics from free text medical records)	**Category 3** Example: Relevant provider information (eg, quality, efficiency or cost of providers/pathways)	**Category 3** Example: Improving existing internal knowledge management systems

LLM, large language model.

Fig 4.4.b Framework for assessing fundamental limitations in LLM applications in healthcare – Table 2. [85]

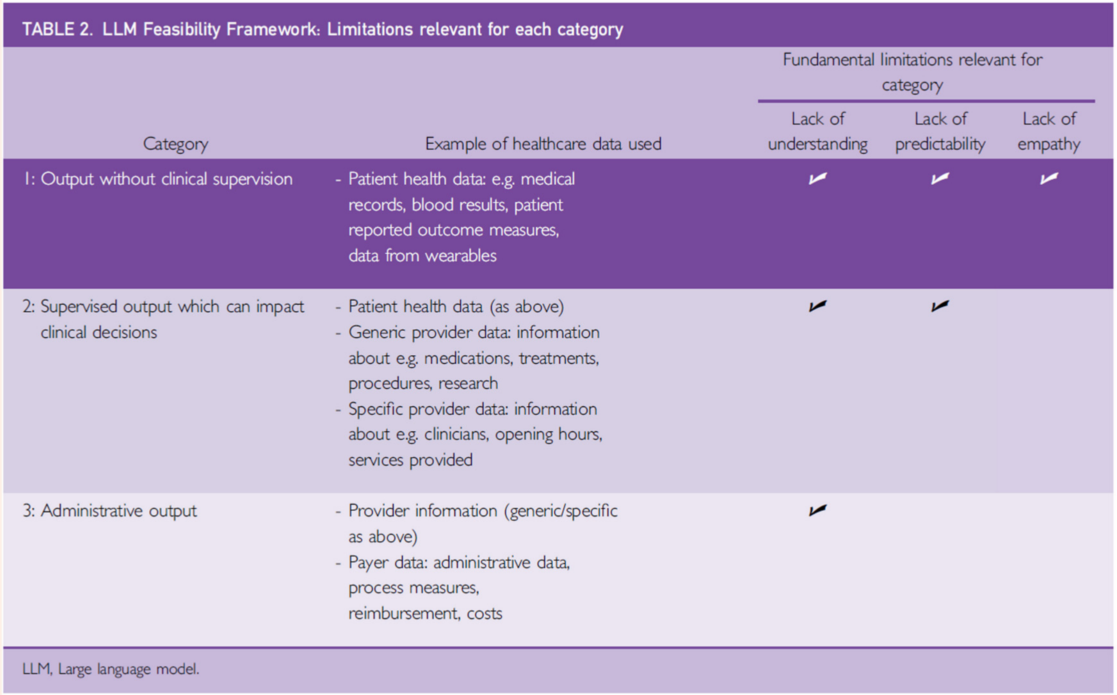


TABLE 2. LLM Feasibility Framework: Limitations relevant for each category

Category	Example of healthcare data used	Fundamental limitations relevant for category		
		Lack of understanding	Lack of predictability	Lack of empathy
1: Output without clinical supervision	- Patient health data: e.g. medical records, blood results, patient reported outcome measures, data from wearables	✓	✓	✓
2: Supervised output which can impact clinical decisions	- Patient health data (as above) - Generic provider data: information about e.g. medications, treatments, procedures, research - Specific provider data: information about e.g. clinicians, opening hours, services provided	✓	✓	
3: Administrative output	- Provider information (generic/specific as above) - Payer data: administrative data, process measures, reimbursement, costs	✓		

LLM, Large language model.

Recommendations

LLMs have the potential to automate and significantly improve three core healthcare activities: communication, applying general knowledge and documentation. Considering the challenges that healthcare is facing and the rapid development we've already seen, LLM applications will become more common and play a more important role in healthcare. However, adopting this technology will require a balancing act from providers as it affects core activities.

In order to reap the benefits of LLMs it will be important for healthcare systems to:

1. **Do the math.** It's easy to write off LLMs as a new hype. However, the time saved by solely automating a majority of documentation is so significant that it warrants serious interest. Calculating an estimated benefit can both clarify why it's worth exploring, as well as guiding the implementation plan and evaluation of investment.

2. **Digitalize healthcare processes** so that LLMs can be applied in an integrated way. It will be challenging to reap benefits from LLM systems without other supporting digital infrastructure. This entails shifting from paper records to EHRs, and ensuring a coherent digital infrastructure to facilitate communication and coordination with patients. Digitalization, done correctly, can independently also free up time.

3. **Increase knowledge about LLMs** in order to identify and mitigate risks. Researchers and developers are moving ahead rapidly, but clinicians and clinical decision-makers need to have a fundamental understanding in order to guide the

development of new knowledge and new LLM applications.

4. **Develop in-house capabilities to guide LLM development and implementation.** As software becomes an increasingly central part of healthcare provision, providers need to be able to control certain activities. Certain technical knowledge is required in order to understand and manage some of LLMs' risks. Some providers will need to develop new capabilities, within for example data science and user experience (UX) in order to be able to navigate this.

5. **Spend time and effort on change management.** Changing any ways of working in healthcare is challenging. Comprehensive clinician education will be needed as the expected benefits from LLMs are contingent on clinicians using a new type of software.

6. **Evaluate investments and spread learnings.** LLM applications are highly contextual and continuously developing. That's why real-life observational evaluations will often be as (or more) relevant for providers to assess the expected effect, compared to

rigorous simulations. The more providers that share their real-life learnings of implementing LLMs, the more other providers can avoid repeating the same mistakes, and instead reap the same successes.

References

Answer to Fig 3.4.1: In both examples the left participant is an AI chatbot and the right participant is a human. Tricky, right?

1. Toscano F, O'Donnell E, Broderick JE, et al. How Physicians Spend Their Work Time: an Ecological Momentary Assessment. *J Gen Intern Med.* 2020;35:3166-3172.
2. Sinsky C, Colligan L, Li L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Ann Intern Med.* 2016;165:753-760.
3. Dubois RW, Dean BB. Evolution of clinical practice guidelines: evidence supporting expanded use of medicines. *Dis Manag.* 2006;9:210-223.
4. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ.* 1996;312:71-72.
5. Doctors (indicator): OECD; 2023.
6. Boniol M, Kunjumen T, Nair TS, Siyam A, Campbell J, Diallo K. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? *BMJ Glob Health.* 2022;7.
7. Ortiz-Ospina E, Roser M. Healthcare Spending: Our World In Data; 2017.
8. UN. World Population Prospects 2022.
9. World Population Prospects 2022. Median age: Our World in Data; 2021.
10. Moynihan RN, Cooke GP, Doust JA, Bero L, Hill S, Glasziou PP. Expanding disease definitions in guidelines and expert panel ties to industry: a cross-sectional study of common conditions in the United States. *PLoS Med.* 2013;10:e1001500.
11. Folkmängden efter region, civilstånd, ålder och kön. År 1968 - 2022: SCB; 2023.
12. Medellivslängden i Sverige: SCB; 2023.
13. Health spending: OECD; 2021.
14. Maas C, van Klaveren D, Visser O, et al. Number of life-years lost at the time of diagnosis and several years post-diagnosis in patients

with solid malignancies: a population-based study in the Netherlands, 1989-2019. *EClinicalMedicine.* 2023;60:101994.

15. Steen Carlsson K, Berne C, Johansen P, Lanne G, Gerdtham U-G. Behandling av diabetes i ett hundraårigt perspektiv: SNS; 2013.
16. Jayachandran S, Lleras-Muney A, Smith, V. K. Modern Medicine and the Twentieth Century Decline in Mortality: Evidence on the Impact of Sulfa Drugs. *American Economic Journal: Applied Economics.* 2010;2:118-146.
17. Armstrong GL, Conn LA, Pinner RW. Trends in infectious disease mortality in the United States during the 20th century. *JAMA.* 1999;281:61-66.
18. Bunker JP. The role of medical care in contributing to health improvements within societies. *Int J Epidemiol.* 2001;30:1260-1263.
19. Crawford J. Draining the swamp - How sanitation fought disease long before vaccines or antibiotics: The Roots of Progress; 2020.
20. Johansson M, Guyatt G, Montori V. Guidelines should consider clinicians' time needed to treat. *BMJ.* 2023;380:e072953.
21. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the Time Needed to Provide Adult Primary Care. *J Gen Intern Med.* 2023;38:147-155.
22. Dalen JE, Ryan KJ, Alpert JS. Where Have the Generalists Gone? They Became Specialists, Then Subspecialists. *Am J Med.* 2017;130:766-768.
23. Janchenko G. The Impact of Electronic Health Record Systems on Physician Productivity. *Issues in Information Systems.* 2020;21:1-8.
24. Howley MJ, Chou EY, Hansen N, Dalrymple PW. The long-term financial impact of electronic health record implementation. *Journal of the American Medical Informatics Association.* 2014;22:443-452.
25. Meyerhoefer CD, Deily ME, Sherer SA, et al. The Consequences of Electronic Health Record Adoption for Physician Productivity and Birth Outcomes. *ILR Review.* 2016;69:860-889.
26. Yarnall KS, Pollak KI, Ostbye T, Krause KM, Michener JL. Primary care: is there enough time for prevention? *Am J Public Health.* 2003;93:635-641.
27. Privett N, Guerrier S. Estimation of the Time Needed to Deliver the 2020 USPSTF Preventive Care Recommendations in Primary Care. *Am J Public Health.* 2021;111:145-149.
28. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions

Posted to a Public Social Media Forum. *JAMA Intern Med.* 2023;183:589-596.

29. CPG Infobase: Clinical Practice Guidelines: CMA; 2023.

30. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ.* 2016;353:i2139.

31. Newman-Toker DE, Nassery N, Schaffer AC, et al. Burden of serious harms from diagnostic error in the USA. *BMJ Qual Saf.* 2023.

32. Centola D, Becker J, Zhang J, Aysola J, Guilbeault D, Khoong E. Experimental evidence for structured information-sharing networks reducing medical errors. *Proc Natl Acad Sci U S A.* 2023;120:e2108290120.

33. Chen CL, McLeod SD, Lietman TM, et al. Preoperative Medical Testing and Falls in Medicare Beneficiaries Awaiting Cataract Surgery. *Ophthalmology.* 2021;128:208-215.

34. Simos D, Catley C, van Walraven C, et al. Imaging for distant metastases in women with early-stage breast cancer: a population-based cohort study. *CMAJ.* 2015;187:E387-397.

35. Morgan DJ, Pineles L, Owczarzak J, et al. Clinician Conceptualization of the Benefits of Treatments for Individual Patients. *JAMA Netw Open.* 2021;4:e2119747.

36. Arkes HR, Aberegg SK, Arpin KA. Analysis of Physicians' Probability Estimates of a Medical Outcome Based on a Sequence of Events. *JAMA Netw Open.* 2022;5:e2218804.

37. Krouss M, Croft L, Morgan DJ. Physician Understanding and Ability to Communicate Harms and Benefits of Common Medical Treatments. *JAMA Intern Med.* 2016;176:1565-1567.

38. Davies K. Evidence-based medicine: is the evidence out there for primary care clinicians? *Health Info Libr J.* 2011;28:285-293.

39. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. *Ann Intern Med.* 1991;114:576-581.

40. Chambliss ML, Conley J. Answering clinical questions. *J Fam Pract.* 1996;43:140-144.

41. Barkman C, Aasa L. Onödig administration i sjukvården. *Forum för Health Policy.* 2023.

42. Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time Spent on Dedicated Patient Care and Documentation Tasks Before and After

the Introduction of a Structured and Standardized Electronic Health Record. *Appl Clin Inform.* 2018;9:46-53.

43. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: A systematic review. *Health Policy.* 2018;122:827-836.

44. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. *J Am Med Inform Assoc.* 2010;17:49-53.

45. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. *JAMA Netw Open.* 2022;5:e2233348.

46. Lauridsen A, Lundqvist L. *Kartläggning av dubbeldokumentation i patientjournalen - förekomst och uppfattningar*: Fakulteten för samhälls- och livsvetenskaper, Karlstads universitet; 2008.

47. Sharp L, Klinga C, Hansson J, Sachs MA. [Electronic health records risk patient safety. Audit of medical records shows serious deficiencies in documentation]. *Lakartidningen.* 2014;111:868-871.

48. Törnqvist J, Törnvall E, Jansson I. Double documentation in electronic health records. *Nordic Journal of Nursing Research.* 2016;36:88-94.

49. Platform24. Independent Evaluation of Efficiency Following Digitalization of Administration Flow2022.

50. Judson TJ, Odisho AY, Young JJ, et al. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. *J Am Med Inform Assoc.* 2020;27:1450-1455.

51. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med.* 2011;104:510-520.

52. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023;620:172-180.

53. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617.* 2023.

54. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. *medRxiv.* 2023:2023.2006.2004.23290939.

55. Cornall J. Life saved: AI discovers existing drug works for rare disease: Labiotech; 2023.

56. Can artificial intelligence restore joy to the practice of medicine? The prognosis is good: Nuance Healthcare; 2023.
57. Ramamurthy R. Summarizing patient histories with GPT-4: Medium; 2023.
58. Quach K. Healthcare org with over 100 clinics uses OpenAI's GPT-4 to write medical records: The Register; 2023.
59. 78% average reduction in documentation time: Ambience Healthcare; 2023.
60. Pimenta D. Medical notes summarisation performance in human clinicians vs LLM: a feasibility study.2023.
61. Jannai D, Meron A, Lenz B, Levine Y, Shoham Y. Human or Not? A Gamified Approach to the Turing Test. *arXiv preprint arXiv:2305.20010.* 2023.
62. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an Upgraded Artificial Intelligence Chatbot for Ophthalmic Knowledge Assessment. *JAMA Ophthalmol.* 2023.
63. Iapoce C. Artificial Intelligence Chatbot Appears to Improve on Ophthalmic Knowledge Assessment: HCP Live; 2023.
64. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol.* 2023;229:172 e171-172 e112.
65. Question Answering on PubMedQA: Papers with Code; 2023.
66. Kiela D, Bartolo M, Nie Y, et al. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337.* 2021.
67. Vardi MY. Artificial intelligence: past and future. *Communications of the ACM.* 2012;55:5-5.
68. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: FDA; 2022.
69. Clarke AC. Clarke's Third Law on UFO's. *Science.* 1968;159:255-255.
70. Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat Neurosci.* 2023;26:858-866.
71. Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv.* 2023:2022.2011.2018.517004.
72. O'Brien SG, Guilhot F, Larson RA, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-

phase chronic myeloid leukemia. *N Engl J Med.* 2003;348:994-1004.

73. Iqbal N, Iqbal N. Imatinib: a breakthrough of targeted therapy in cancer. *Chemother Res Pract.* 2014;2014:357027.

74. Ball P. The lightning-fast quest for COVID vaccines - and what it means for other diseases. *Nature.* 2021;589:16-18.

75. Venkitachalam L, Kip KE, Selzer F, et al. Twenty-year evolution of percutaneous coronary intervention and its impact on clinical outcomes: a report from the National Heart, Lung, and Blood Institute-sponsored, multicenter 1985-1986 PTCA and 1997-2006 Dynamic Registries. *Circ Cardiovasc Interv.* 2009;2:6-13.

76. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. *arXiv.* 2022:2202.03629.

77. Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence.* 2022;4:258-268.

78. Feng S, Park CY, Liu Y, Tsvetkov Y. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *arXiv preprint arXiv:2305.08283.* 2023.

79. Angwin J, Larson J. Machine Bias: ProPublica; 2016.

80. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc.* 2022;37:1353-1359.

81. Morris R., Kouddous K., Kshirsagar R., Schueller S. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *J Med Internet Res.* 2018;20:e10148.

82. Jackson JC, Yam KC, Tang PM, Liu T, Shariff A. Exposure to robot preachers undermines religious commitment. *J Exp Psychol Gen.* 2023.

83. SMER. Kort om Artificiell intelligens i hälso- och sjukvården. 2020;2:1-16.

84. SMER. Konferensrapport Artificiell intelligens – löftesrik teknik med etiska utmaningar. 2019;2:11-14.

85. Ilicki J. A Framework for Critically Assessing ChatGPT and Other Large Language Artificial Intelligence Model Applications in Health Care. *Mayo Clinic Proceedings: Digital Health.* 2023;1:185-188.