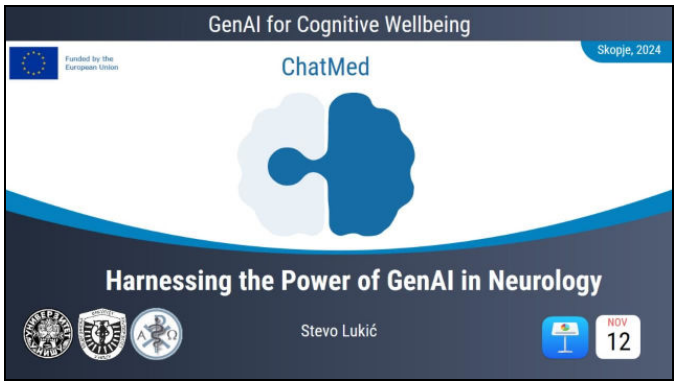
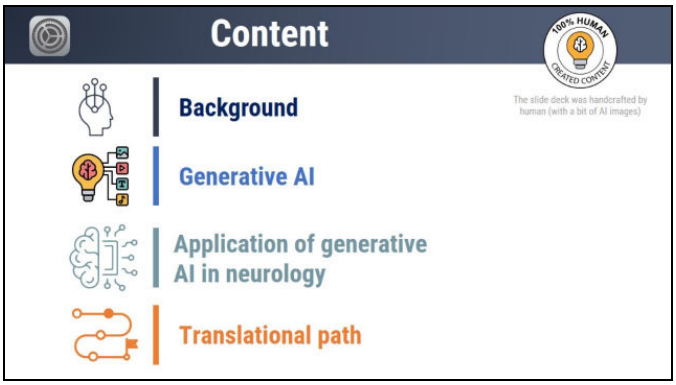
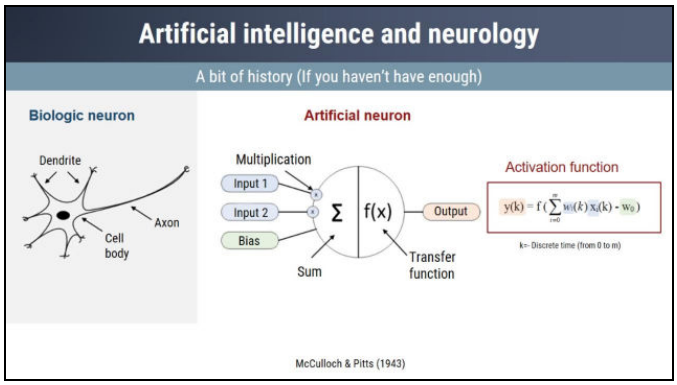

GenAI for Cognitive Wellbeing
Funded by the European Union
Skopje, 2024
ChatMed
Harnessing the Power of GenAI in Neurology
Stevo Lukić
NOV
12

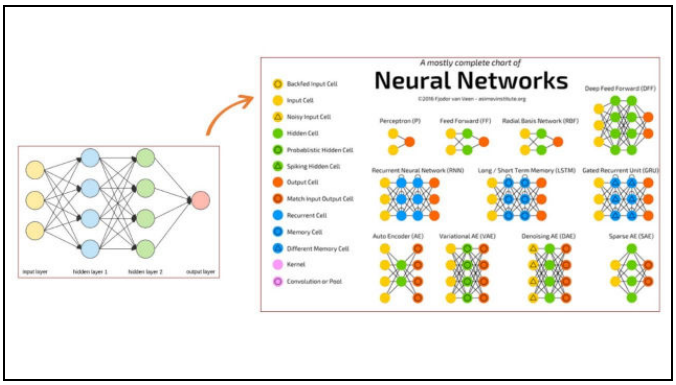
1


Content
100% HUMAN CREATED CONTENT
The slide deck was handcrafted by human (with a bit of AI images)
Background
Generative AI
Application of generative AI in neurology
Translational path

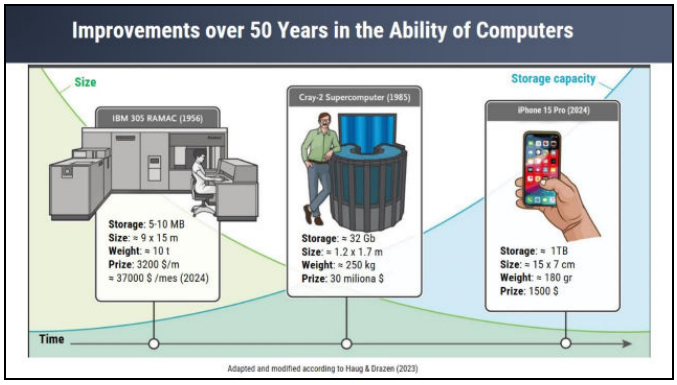
2


Artificial intelligence and neurology
A bit of history (If you haven't have enough)
Biologic neuron
Dendrite
Axon
Cell body
Artificial neuron
Multiplication
Input 1
Input 2
Bias
Σ
f(x)
Output
Sum
Transfer function
Activation function
k– Discrete time (from 0 to m)
McCulloch & Pitts (1943)

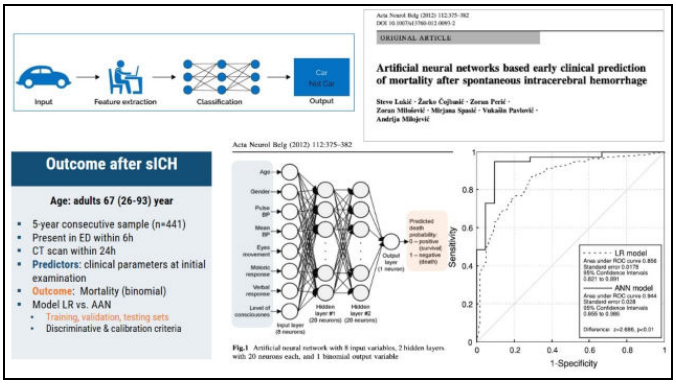
3


A mostly complete chart of
Neural Networks

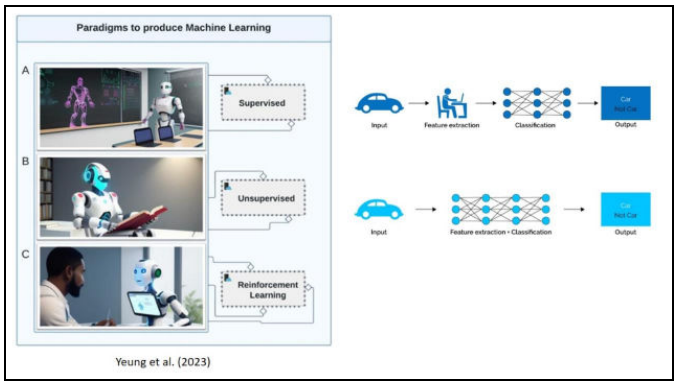
4


Improvements over 50 Years in the Ability of Computers
Size
Storage capacity
IBM 305 RAMAC (1956)
Storage: 5-10 MB
Size: ≈ 9 x 15 m
Weight: ≈ 10 t
Prize: 3200 $/m
≈ 37000 $ /mes (2024)
Cray-2 Supercomputer (1985)
Storage: ≈ 32 Gb
Size: ≈ 1.2 x 1.7 m
Weight: ≈ 250 kg
Prize: 30 miliona $
iPhone 15 Pro (2024)
Storage: ≈ 1TB
Size: ≈ 15 x 7 cm
Weight: ≈ 180 gr
Prize: 1500 $
Time
Adapted and modified according to Haug & Drazen (2023)

5

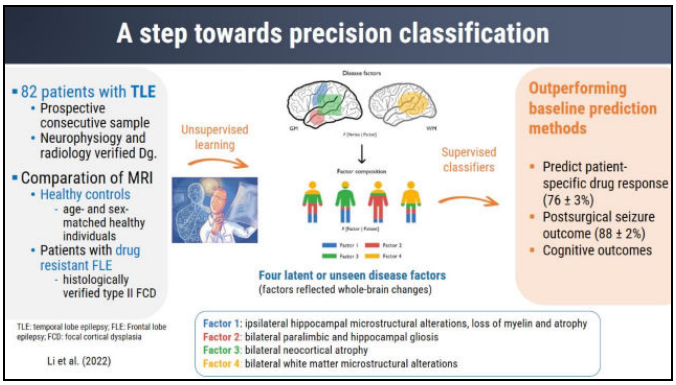

Input
Feature extraction
Classification
Output
Car
Not Car
ORIGINAL ARTICLE
Artificial neural networks based early clinical prediction of mortality after spontaneous intracerebral hemorrhage
Outcome after sICH
Age: adults 67 (26-93) year
5-year consecutive sample (n=441)
Present in ED within 6h
CT scan within 24h
Predictors: clinical parameters at initial examination
Outcome: Mortality (binomial)
Model LR vs. AAN
Training, validation, testing sets
Discriminative & calibration criteria
Sensitivity
1-Specificity
LR model
ANN model

6

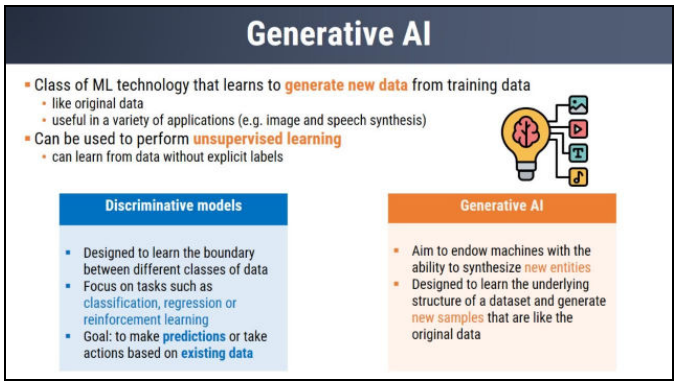


Yeung et al. (2023)

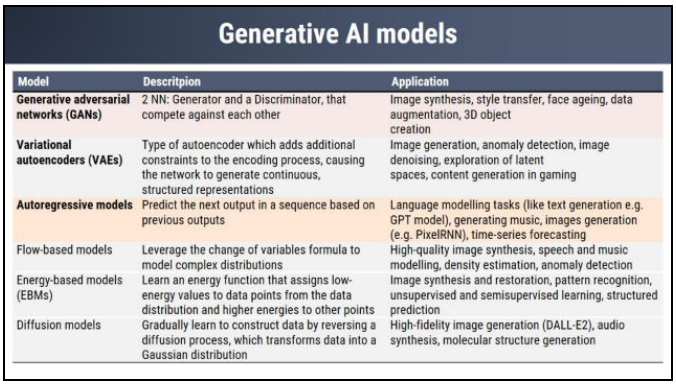
7



8

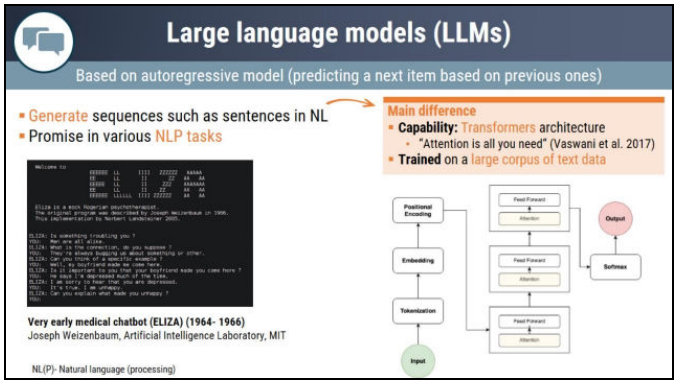


9



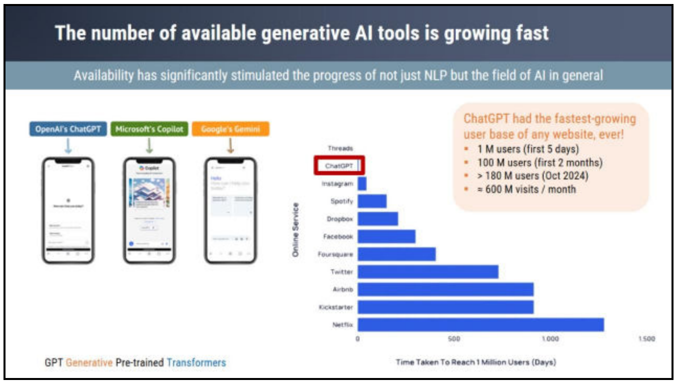
Model	Descritpion	Application
Generative adversarial networks (GANs)	2 NN: Generator and a Discriminator, that compete against each other	Image synthesis, style transfer, face ageing, data augmentation, 3D object creation
Variational autoencoders (VAEs)	Type of autoencoder which adds additional constraints to the encoding process, causing the network to generate continuous, structured representations	Image generation, anomaly detection, image denoising, exploration of latent spaces, content generation in gaming
Autoregressive models	Predict the next output in a sequence based on previous outputs	Language modelling tasks (like text generation e.g. GPT model), generating music, images generation (e.g. PixelRNN), time-series forecasting
Flow-based models	Leverage the change of variables formula to model complex distributions	High-quality image synthesis, speech and music modelling, density estimation, anomaly detection
Energy-based models (EBMs)	Learn an energy function that assigns low-energy values to data points from the data distribution and higher energies to other points	Image synthesis and restoration, pattern recognition, unsupervised and semisupervised learning, structured prediction
Diffusion models	Gradually learn to construct data by reversing a diffusion process, which transforms data into a Gaussian distribution	High-fidelity image generation (DALL-E2), audio synthesis, molecular structure generation

10

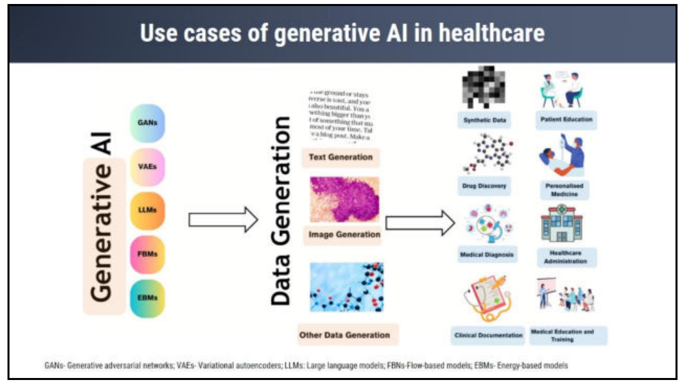


11

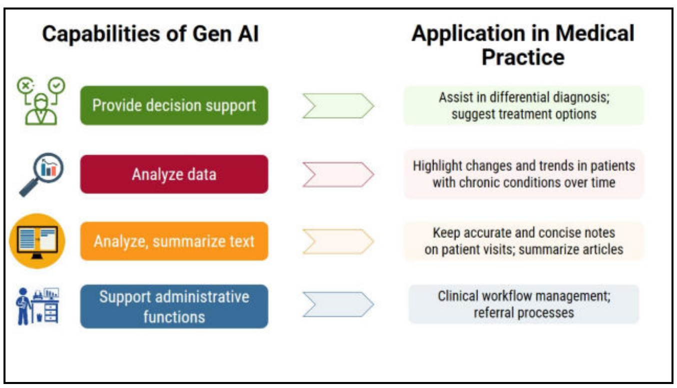
12



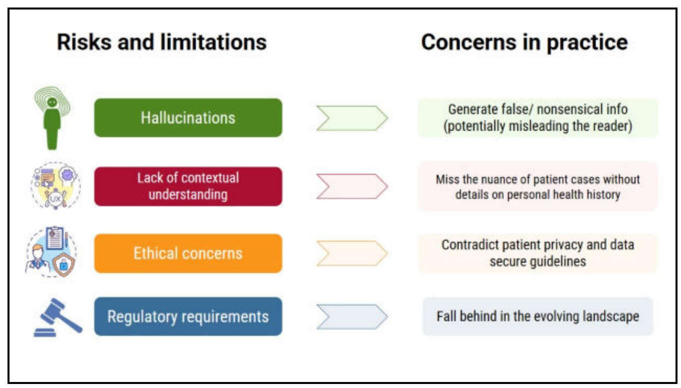
13



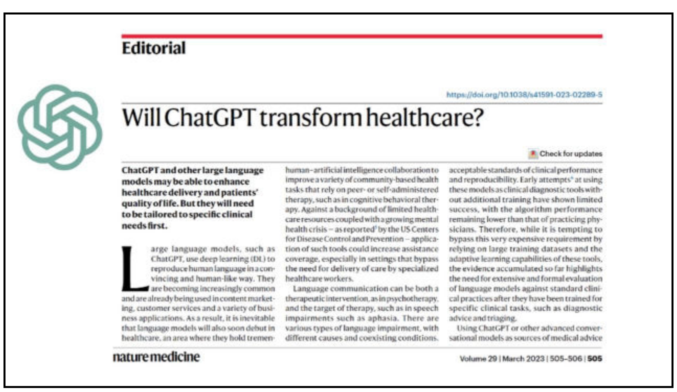
14



15



16



Editorial

https://doi.org/10.1038/s41591-023-02289-5

Will ChatGPT transform healthcare?

Check for updates

ChatGPT and other large language models may be able to enhance healthcare delivery and patients' quality of life. But they will need to be tailored to specific clinical needs first.

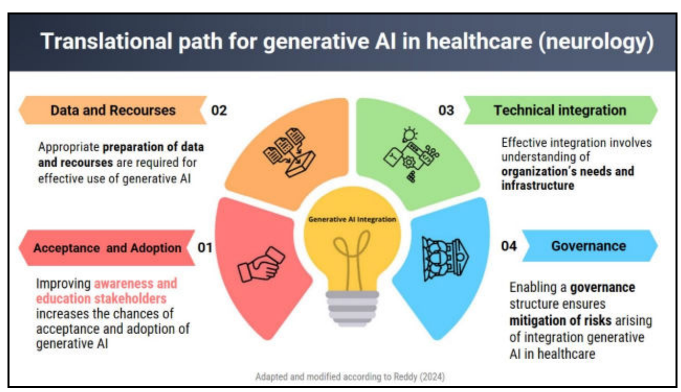
Large language models, such as ChatGPT, use deep learning (DL) to reproduce human language in a convincing and human-like way. They are becoming increasingly common and are already being used in content marketing, customer services and a variety of business applications. As a result, it is inevitable that language models will also soon debut in healthcare, an area where they hold tremen- human–artificial intelligence collaboration to improve a variety of community-based health tasks that rely on peer- or self-administered therapy, such as in cognitive behavioral therapy. Against a background of limited healthcare resources coupled with a growing mental health crisis – as reported[3] by the US Centers for Disease Control and Prevention – application of such tools could increase assistance coverage, especially in settings that bypass the need for delivery of care by specialized healthcare workers.

Language communication can be both a therapeutic intervention, as in psychotherapy, and the target of therapy, such as in speech impairments such as aphasia. There are various types of language impairment, with different causes and coexisting conditions. acceptable standards of clinical performance and reproducibility. Early attempts[4] at using these models as clinical diagnostic tools without additional training have shown limited success, with the algorithm performance remaining lower than that of practicing physicians. Therefore, while it is tempting to bypass this very expensive requirement by relying on large training datasets and the adaptive learning capabilities of these tools, the evidence accumulated so far highlights the need for extensive and formal evaluation of language models against standard clinical practices after they have been trained for specific clinical tasks, such as diagnostic advice and triaging.

Using ChatGPT or other advanced conversational models as sources of medical advice

nature medicine Volume 29 | March 2023 | 505–506 | 505

17



18

Acceptance and adoption of end user

Fosters trust in AI systems by professionals and patients → Effective use and aids in navigating ethical and regulatory challenges

- Perceived **usefulness**
- Perceived **ease of use**
- Attitude towards using
- Behavioral intention to use
- Actual system use

Investment in improving awareness for all partners is crucial to ensure the effective adoption and utilization of AI in healthcare (neurology)

19

Usefulness of generative AI in medicine

Pro

- Strong performance on medical licensing exams [1]
- Promise in improving ED triage accuracy [2]
- Diagnostic accuracy can reach the accuracy of other experts and surpasses GPs [3]
  - e.g. capabilities of in the localization of acute stroke lesions [4]
- Chatbot generated quality and empathetic responses to patient questions posted in an online forum [5]

Contra

- "Chatbot GPT can be grossly inaccurate" [6]
- "AI-Generated clinical summaries require more than accuracy" [7]
- Fail to assess many skills necessary for deployment in a realistic clinical decision-making environment [8]
  - including gathering information
  - adhering to guidelines
  - integrating into clinical workflows
- Did not significantly improve clinical reasoning compared with conventional resources in RCT [9]

[1] Kung et al. (2023); [2] Kaboudi et al. (2024); [3] Nógrádi et al. (2024); [4] Lee et al. (2024); [5] Auser et al. (2023)
[6] Diamandis (2023); [7] Goldman et al. (2024); [8] Hager et al. (2024); [9] Goh et al. (2024)

20

LLM influence on diagnostic reasoning?

RCTs are considered the gold standard in clinical research for evaluating the **efficacy of innovations**

POPULATION
26 Attending physicians
24 Resident physicians
US-trained physicians with training in family medicine, internal medicine, or emergency medicine

INTERVENTION
50 Participants randomized
25 Generative artificial intelligence (AI) chatbot
Participants with access to AI chatbot were allocated 60 min to review up to 6 clinical vignettes
25 Conventional resources
Participants with access to conventional online resources such were allocated 60 min to review up to 6 clinical vignettes

FINDINGS
There was no significant difference in diagnostic performance between physicians in the chatbot group and the conventional resource group
Median diagnostic reasoning score
AI chatbot 76%
Conventional resources 74%
Adjusted difference between groups:
2 percentage points (95% CI, −4 to 8 percentage points); P = .60

SETTINGS / LOCATIONS
3 Academic medical centers in the US

PRIMARY OUTCOME
Diagnostic performance, assessed using a standardized rubric including differential diagnosis accuracy, appropriateness of supporting/opposing factors, and next evaluation steps; graded via blinded expert consensus

Stanford MEDICINE

Goeth et al. JAMA Network Open. 2024;7(10):e2440969

RCT- Randomized controlled trials

21

1. Articulate the specific problems or use cases that would benefit from the incorporation of AI
2. Mirrors the tried-and-true practice adopted in the development of drug therapies and medical devices, which begins with the clinical indication

CLINICAL INDICATION	EXAMPLE OF AI
Interaction with patients	Ambient voice dictation, scheduling, HER inbox tool
Risk stratification (precision medic.)	Patient risk assessment tools
Diagnosis	Analysis of clinical data or imaging (eg, EEG)
Interpretation of laboratory results	Analysis and description of test results
Eliciting patient preferences or behavior changes	Conventional chatbot
Procedures	Surgical assistance
Prescribing medication	Drug interaction assesment
Patient or population monitoring	Glu monitoring, population- at risk-monitoring
Research and learning	Research participant identification and egagement
Continuing education and training	Virtual reality case simulation

Patel et al. JAMA (2024)

22

Development of clinical AI analogous to pharmacotherapies (testing and adoption)

	DRUG DEVOPMENT	AI DEVOPMENT
Phase 1	Pharmacokinetics and pharmacodynamics studies	**Model development:** Testing on large retrospective dataset
Phase 2	Collection of safety data and efficacy signals	**Silent mode evaluation:** Prospective performance evaluation in the clinical environment without notification to clinical teams
Phase 3	Clinical trail with outcomes assessment	**Clinical deployment:** AI prospectively implement wit measurement of downstream care and outcomes (consider RCT design for high-risk clinical indication)
Phase 4	Real- world use (postmarket registry of use in practice)	**Monitoring and assurance network:** Deployment in multiple environments with local monitoring of performance and outcome

Indication and risk of the AI technology would match the methodology for clinical testing and adoption
- Low vs. high-risk AI tools

Patel et al. JAMA (2024)

23

Evidence for health AI technologies

Should be evaluated and reviewed like all clinical practice guidelines

SIZE OF TREATMENT EFFECT	Class I	Class IIa	Class IIb	Class III
	Benefit >>> risk	Benefit >> risk	Benefit > risk	No benefit or harm
	Should be performed	Reasonable to performed	May be considered	Should not be performed

LEVEL OF EVIDENCE	Level A	Level B	Level C
	Multiple population	Limited population	Very limited data
	Data derived from multiple RCTs or meta-analysis	Data derived from single RCTs or nonrandomized studies	Expert consensus of opinion, case studies, or standard of care

Patel et al. JAMA (2024)

24

Bedi et al. Nat Med. 2024;30(9):2409-10.

Machine learning

Evaluating the clinical benefits of LLMs

Suhana Bedi, Sneha S. Jain & Nigam H. Shah

Nature Medicine 30, 2409–2410 (2024) | Cite this article

...

Although large language models (LLMs) show promise in controlled settings, a study now exposes their limitations in real-world clinical applications and **points the way towards robust evaluation and benchmarking before clinical use**

Bedi et al. *JAMA*. Published online October 15, 2024.

JAMA | Original Investigation | AI IN MEDICINE

Testing and Evaluation of Health Care Applications of Large Language Models

A Systematic Review

CONCLUSIONS AND RELEVANCE

- Existing evaluations of LLMs mostly focus on accuracy of question answering for medical examinations, without consideration of real patient care data.
- Dimensions such as fairness, bias, and toxicity and deployment considerations received limited attention.
- **Future evaluations should adopt standardized applications and metrics**, use clinical data, and broaden focus to include a wider range of tasks and specialties.

25

Evaluation of genAI in neurology (1)

Dimension of Evaluation	Definition	Metric Examples
Accuracy	Measures how close the LLM output is to the true or expected answer	Human evaluated correctness, ROUGE, MEDCON
Calibration and Uncertainty	Measures how uncertain or underconfident an LLM is about its output for a specific task	Human evaluated uncertainty, calibration error, Platt scaled calibration slope
Robustness	Measures the LLMs resilience against adversarial attacks and perturbations like typos	Human evaluated robustness, exact match on LLM input with intentional typos, F1 on LLM input with intentional use of word synonyms

Bedi S et al. MedRvix August 16. 2024.

26

Evaluation of genAI in neurology (2)

Dimension of Evaluation	Definition	Metric Examples
Comprehensiveness	Measures how well an LLMs output coherently and concisely addresses all aspects of the task and reference provided	Human evaluated comprehensiveness, fluency, UniEval relevance
Fairness, bias and toxicity	Measures whether an LLMs output is equitable, impartial, and free from harmful stereotypes or biases, ensuring it does not perpetuate injustice or toxicity across diverse groups	Human evaluated toxicity, counterfactual fairness, performance disparities across race
Deployment considerations	Measures the technical and parametric details of an LLM to generate a desired output	Cost, latency, inference runtime

Bedi et al. MedRvix (August 16. 2024.)

27

Will generative AI replace doctors?

Fears of job automation and resultant unemployment are unfounded

Core activities:

- Diagnostic and multimodal interpretation for patient care
- Clinical decision-making
- Procedural tasks
- Human communication

Many related activities with subtasks

- Administration
- Para-clinical activities
- Research and education

Different combinations of cognitive, communication and motor contributions, which are not easily segregable

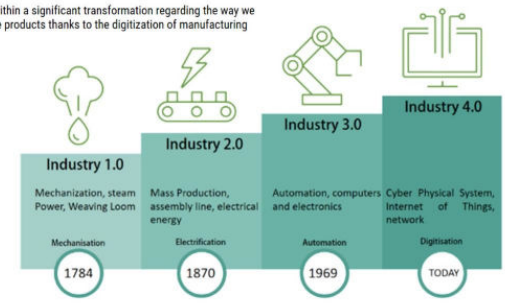
28

Fountain of Creativity or Pandora's Box?

- The practice of medicine is much more than just processing information and associating words with concepts
- AI and machine learning will not put health professionals out of business
- Rather, they will make it possible for health professionals to do their jobs better and leave time for the human-human interactions that make medicine the rewarding profession we all value

29

- We're within a significant transformation regarding the way we produce products thanks to the digitization of manufacturing

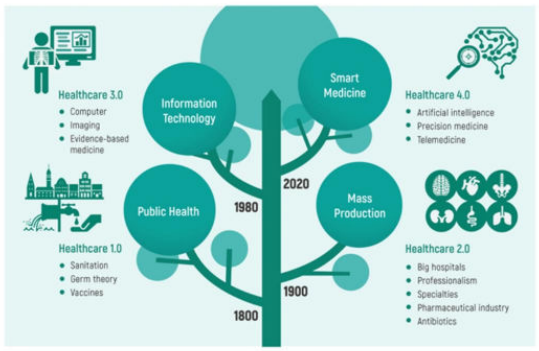


- This transition is so compelling that it is being called **Industry 4.0** to represent the fourth revolution that has occurred in manufacturing

30

- Technology is rapidly changing the way medicine is practiced

- **Medicine 4.0** is a term used to denote a new era in medicine, based on the use of information and communication technologies to improve health care



31

Conclusions

- Generative AI is proving to be a change catalyst across various industries, and the healthcare sector is no exception
- Despite the potential benefits, application of generative AI in healthcare raises some concerns
- Technological progress alone will not revolutionize healthcare overnight
- Guided by wisdom and compassion, generative AI may help get closer to health- care ideals (so many now lack): quality, accessibility and humane care for all

32