Approaches to Evaluating Large Language Models and Conversational Agents for Healthcare Applications.

Arkangel AI[a]

[a]Natalia Castano-Villegas [2] , Isabella Llano[1] , Julián Martinez[3], Daniel Jimenez[3], María Camila Villa[3], Jose Zea[4], Laura Velasquez[4]

1. Medical Epidemiologist at Arkangel AI
2. Biomedical Researcher Engineer at Arkangel AI
3. ML Engineer at Arkangel AI
4. Co-Founder at Arkangel AI

ABSTRACT

Large language models and conversational agents are being widely introduced to healthcare, given their powerful capabilities in knowledge acquisition, instruction comprehension, generalization, planning, reasoning, and their ability to interact effectively with humans. The landscape is continuously evolving, and tools are becoming increasingly complex. Current evaluation strategies include question-answering datasets based chiefly on clinical boarding exams that assess the model's specific knowledge through multiple-choice and open-answer questions and human evaluation in which qualified individuals appraise the accuracy, quality, and correctness of the algorithm responses. However, researchers and clinicians have not agreed on a "gold standard" for evaluation to ensure models' safety, accuracy and efficacy in clinical settings. We aim to present an overview of the current options for assessment, their challenges, and new ways to face them.

INTRODUCTION

STATE-OF-THE-ART VALIDATION STRATEGIES FOR LARGE LANGUAGE MODELS IN HEALTHCARE APPLICATIONS

Language is a system of knowledge (1) which broadly defines us as humans. From the caves to ancient Greece to Egyptian hieroglyphics and the French Declaration of Human Rights, language is one of the characteristics that has allowed our race to master knowledge and power over other species. It was defined as "the medium through which one conveys complex thought using arbitrary symbols in a significant syntax"(2). Although not exclusive to *Homo Sapiens*, the thirst

for more (food, knowledge, money, power, etc.) has driven our race's progress and evolution. Language helped our ancestors to communicate and pass on survival skills. It also made it possible to establish and maintain traditions and unity between tribes and families. It helped to spread knowledge and recent discoveries and made them into the public domain.

The deciphering of hieroglyphics, symbols carved in stone or imprinted in papyrus, used by the Egyptians to represent sounds and concepts (3), was attributed to Jean-François Champollion, a French scholar in the nineteenth century who was able to decode the inscriptions from the Rosetta Stone, comparing the three different scripts on it and finding common patterns which allowed him to determine the meaning of unknown words (the hieroglyphics) from scripts in known ancient Greek (4).

Humans have steered unprecedented scientific and technological advances ever since, culminating with the materialization of Alan Turing's dream in the 1950s. The mathematician, philosopher, computer scientist and cryptanalyst had a forward mentality that was violently reprimanded for the age he lived in, as he was regrettably condemned for being a homosexual, despite his brilliantness and immense contribution to the British Crown during World War II (5). He was the first to describe the capability of machines to learn to perform tasks traditionally carried out by humans, simulating thought processes which demonstrate the ability to teach themselves after initial instruction (6). This concept was widely revindicated in the twenty-first century as Artificial Intelligence (AI) or automatic learning.

Much like the deciphering of the Rosetta Stone, humans learned to use computational advances and power to understand further how some phenomena occur by decrypting their unknown patterns and behaviours based on ground truths accepted by the scientific community. This is how Automatic or Machine Learning (ML) can predict different outcomes in several fields, recognize specific characteristics in diagnostic images in medicine, support decision-making across various professions, and facilitate repetitive processes in general (7). Deep Learning (DL) is a type of ML specialized in processing non-tabular data (ML's speciality). It rather works better with unstructured information, often in the form of Natural (or human) Language (8).

Natural language processing (NLP) is part of ML, which tries to achieve computational analysis of human language on a scale capable of imitating our own abilities. After years of development, Large Language Models (LLMs) have broadened their capacities and are based on massive datasets and billions of hyperparameters to recognize, interpret and generate text effectively (9). The use of LLMs has increased exponentially in the past years due to their powerful capabilities in knowledge acquisition, instruction comprehension, generalization, planning, and reasoning, as well as their ability to interact effectively with humans (10). Conversational Agents (CAs) are LLM-powered software programs that imitate human conversations. They have become popular

in personal and professional settings (11). Notable examples are ChatGPT from Open AI, Claude from Anthropic, and Gemini from Google.

Generative AI is the general term used for highly specialized computational processes. One of its current uses is to train machines to learn conversational patterns to complete natural language tasks, They are often created on demand, including LLMs and CAs (9). The first models were developed and trained using millions of novels, books and document databases. They also included webpages and Wikipedia (12). ChatGPT, one of the most prominent examples, has been improved to the maximum performance described for language processing in its fourth version. It now allows for using multimodal input, such as images and sound.

The landscape of LLMs and CAs is continually and rapidly evolving, permeating more fields and areas of knowledge. These AI tools are being widely developed for the healthcare industry due to their capacity to handle, organize and provide physicians with precise and immediate information retrieved from trusted medical libraries, worldwide updated clinical guidelines and peer-reviewed literature with the highest levels of evidence. However, there are significant challenges to ensuring their safety, validity, and usefulness in real-world health scenarios.

Validation of LLMs as assistant tools in the healthcare industry has proven difficult. First, in a perfect scenario, physicians would evaluate the accuracy and performance of the models' answers to queries. Furthermore, people with clinical training should be the ones who generate the statements and questions designed to train the algorithms. That is the outlook of Human Evaluation (13). Yet, the tremendous amount of data capable of being processed with AI makes this a task that could easily overpower human capacity and would be a lengthy and costly requirement for model deployment (14). Consequently, researchers and developers have approached this by creating information databases in text formats, mainly questions and answers, referred to as Question Answering (QA) Evaluation. The information in QA datasets is used to "test" the conversational model with general or field-specific questions to simulate how humans would be tested (15).

The evaluation and validation of LLMs have involved several methods to test Natural Language Understanding (NLU) with tasks such as sentence completion, text classification, and sentiment analysis (16). Automated metrics like BLEU and ROUGE evaluate models' ability to generate perfect word matches. BLEU was initially developed to assess machine translation tasks; in principle, it compares word matches between candidate translations and reference translations to judge a good from a bad one (17). Similarly, ROUGE was intended to determine a summary's quality by comparing it to an ideal, human-created one. This was achieved by counting the overlapping word sequences and pairs between both summaries (18). The two metrics make a direct word pair comparison of the computer-generated text to human-generated candidates. Nevertheless, these have been criticized for correlating poorly with human judgement. They fall

short on more complex aspects of language generation, such as preserving grammaticality and meaning where sentence splitting is involved (19).

For both human and QA evaluations (evals), some studies have shown that LLMs' output quality depends on the specificity of the input instruction, a characteristic known as prompting (16). Additionally, CAs and LLMs have been known to hallucinate, generating well-written content that appears factual but that is, in fact, unfounded. They can also misinterpret inputs or misalign information to fit a given input (14) and follow the lead of biased information sources. Thus, medical applications using LLMs are still far from autonomous and require robust validations to ensure patient safety and ethical practices.

Researchers must address the need for standardized evaluation strategies for LLMs in healthcare applications since the absence of a "Gold Standard" makes validating models challenging and subjective. Considering the relevance, fast development, and introduction of specialized CAs in medical fields, we have recognised that critical appraisal guidelines to assess LLMs' quality are missing. We aim to summarize the current options explored in LLM and CA evaluation in healthcare and the most significant challenges they currently face.

Methods

We performed an unstructured literature review. Two researchers thoroughly revised forty manuscript methodologies. Study designs included systematic reviews, literature reviews, report papers, expert consensus papers, review and editorial articles, and articles published from 2015 to 2024. Sources consulted were PubMed, Arxiv, and Medrxiv. We also used web leaderboard pages where CAs ranking is displayed by the ML engineer community in real time and the Google Scholar engine to include grey literature and conference book publications.
Additionally, we included insights on current research discussed in the Global Health Symposium in Cali, Colombia, in July 2024, the XXI Accion for Health Congress, organized with the support of the Colombian Government, in Cartagena in August 2024, and the XXVI National Nephrology Congress, organized by the Colombian Association of Nephrology, 2024. Thus, selection criteria were broad, as we did not want to exclude any papers related to LLMs validation strategies, considering how new evidence is being produced every day, and its quality does not necessarily depend on the extended submission and acceptance processes by peer reviewers and high publishing fees. Keywords used were Large Language Models, Conversational Agents, Validity, and Health Care Evaluation.

Our main objective was to describe strategies used in the current literature about LLMs for evaluating the validity of conversational models and present their benefits and improvement opportunities as we build more robust evaluation methods for conversational models in development.

1. QUESTION-ANSWERING EVALUATION

LLMs and CAs are frequently evaluated using question-answering (QA) datasets. A broad variety of datasets containing clinical questions have been created for this purpose.

The MedQA dataset is a clinical question-answer database with multiple-choice answers. It was compiled from medical board questionnaires from the United States (USMLE), Mainland China and Taiwan using three languages: English, simplified Chinese and traditional Chinese. Respectively, each set contains 12723, 34251 and 14123 questions. When introduced in September 2020, the database was used to test state-of-the-art (SoTA) models. At the time, the highest performance scores, measured using accuracy, were for the BioBERTLarge model, with 36.7% and 42.0% for the US and Taiwan datasets and 70.1% for the Mainland China one (15). This model is the fine-tuned version of the BERT (Bidirectional Encoder Representations from Transformers) model, introduced in 2019 (20), which was improved with the bio-medical literature from PubMed (21). The current SoTA reported for MedQA-US is 90.2% accuracy, using GPT-4 and a prompting strategy called Medprompt (22). Med-PaLM 2, another high-performing LLM, achieved an accuracy of 85.4% (23). Finally, a simulacrum methodology for LLM-powered CAs named MedAgent-Zero applied to GPT-4 managed to reach 93.06% accuracy on a subset of the dataset, covering major respiratory diseases (24). The human passing score on the USMLE is 60%, and the human expert score is 87% (25).

MedMCQA is another large-scale multiple-choice question-answering dataset that compiles Indian medical entrance exam questions. It includes over 194000 multiple choice questions from the All-India Institute of Medical Science (AIIMS) and the National Eligibility Entrance Test for Postgraduation (NEET PG). It covers questions on 21 medical fields and uses external knowledge sources such as context (Wikipedia and PubMed) and internal knowledge (the model itself), which helps calculate the contribution of other resources to its output. At the time of introduction, in May 2022, SoTA was at 41% accuracy with the internal approach and 47% with the external approach, using the model PubMedBERT (26). Currently, SoTA is at 72.3% accuracy for the model Med-PaLM 2 (23). The human passing score was 50%, while the expert score was 90% (25).

PubMedQA, introduced in September 2019, was designed to assess medical reading comprehension on medical abstracts from PubMed. It contains medical research questions that can be answered with the options yes, no, or maybe, using their corresponding abstracts, except for the conclusions. It includes 1000 expert-annotated instances, 61200 unlabeled and 211300 artificially generated. The initial performance was 68.1% accuracy, using fine-tuned BioBERT

(27). As with MedQA, SoTA for this dataset is 81.6% accuracy using GPT-4 with Medprompt (22). The human expert score for this dataset has been reported as 78% (25).

The Measuring Massive Multitask Language Understanding (MMLU) dataset is a benchmark constructed to assess models across a diverse set of topics spanning 57 different subjects, including STEM, humanities, social sciences, and medical sciences. The dataset spans multiple choice questions that range in difficulty from elementary to professional levels. The questions were collected from open-access sources, including practice questions for licensing examinations like the USMLE. Each subject's test set contains a minimum of 100 questions. The MMLU clinical topics subset includes the subjects of clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine (28). This dataset was evaluated with Med-PaLM 2 on clinical knowledge, medical genetics, anatomy, professional medicine, and college biology. The accuracies were 88.7%, 92.0%, 84.4%, 95.2%, and 95.8%, respectively.

MeDiaQA, constructed from natural telemedicine dialogues from China's most common online medical consultation websites (haodf.com and dxy.com), includes 22k multiple-choice questions from over 11k dialogues between patients and doctors covering 150 specialities. This dataset evaluates the model's ability to comprehend the interaction and clinical content of conversations between physicians and patients and answer questions based on them. Tasks for assessing LLMs include extracting evidence from the dialogue reasoning and performing clinical mathematical calculations such as the medication the patient should take (8).

Another language-specific dataset is FrenchMedMCQA, which comprises 3105 questions from French medical exams. This dataset includes multiple-choice questions with both single and multiple answers. It is evaluated on overall accuracy (Hamming score) and exact match ratio (EMR), which dictates that multiple answers must all be correct to count as correct. Metrics achieved with the model CamemBERT were a 36.24% Hamming score and 16.55% EMR (29)**.** Similarly, the Chinese medical QA dataset Huatuo-26M has 26 million QA pairs collected from online medical consultation websites, medical encyclopedias, and medical knowledge bases. The benchmark presented for this dataset comes from fine-tuned versions of the models T5 and GPT-2, obtaining a maximum performance of 33.21% and 30.48% on the metric ROUGE (30).

MultiMedQA is a benchmark of six medical QA datasets: MedQA, MedMCQA, PubMedQA, LiveQA, MedicationQA and MMLU clinical topics, and they also introduce a seventh dataset, HealthSearchQA, consisting of 3173 commonly searched consumer online medical questions. These datasets vary in: format, some are multiple-choice, and others are long-answer; the capabilities they assess, from medical reasoning to recall of facts; in their sources for clinical questions, which include professional medical exams, medical research or consumer questions and in the presence of labels and/or explanations (31).

Limitations

LLMs are influenced by the specific instructions given. Liévin et al. evaluated models using direct and chain of thought (CoT) prompting strategies. For the first one, the model answered "The answer is" to each question. For CoT, the model had a reasoning prompt, such as "Let's think step by step", and an extractive prompt, "Therefore, the answer is". According to the authors, CoT prompting allowed them to conduct a deeper analysis of the model's responses. However, after evaluating 50 answers, 62% included at least one incorrect reasoning step, 58% had inaccurate or insufficient knowledge, and 36% had incorrect reading comprehension. They also found biases induced by both prompting strategies, leading towards one answer (25). This exemplifies the unpredictable nature of prompting evaluations that are useful for interpreting the answers but don't necessarily provide clear associations.

Another factor that could influence LLMs was evaluated by Reichenpfader et al. They introduced an evaluation framework that emulates populations with different levels of health literacy to assess a model's answers. These prompts were generated by another LLM (OpenAI's GPT-4) to automate question generation and evaluation. This framework obtained a lower accuracy at higher literacy levels. However, they report that high literacy-generated questions often included long and complicated word variations that proved challenging for the model (32). This approach was only tested on a set of predefined mammography questions, limiting the conclusions that can be made of its evaluation utility. However, it presents an interesting approach and highlights the importance of building evaluations to suit the intended user's health literacy. It also underscores the limitations of automated evaluation, where models can generate assessments that poorly represent reality by using overly complex formulations and unnatural word combinations.

Additionally, language models are meant to answer queries independently without choosing from a set of options. That is why QA evaluations cannot fully represent the intended use. QA datasets can only test the knowledge and understanding of the model of a particular question. Still, they do not test how the model communicates with the user, which is crucial for CAs and their adoption in clinical settings. Also, the agent cannot identify when diverse user input exists. Evaluations using multiple prompts for the same question have been developed in response.

As research advances, fine-tuned models for specific areas or tasks are being developed using domain-specific information to create specialized models whose evaluations rely on precise knowledge scales. While it is possible to find adequately validated scales for many specialties and/or diseases, this is only the case for some of them, and general standards of reference still need to be improved.

2. HUMAN EVALUATION

Some studies have proposed different human evaluation (or human eval, HE) frameworks to assess quality beyond automatic response, mainly since LLMs used in medicine put them directly in the hotspot for human interaction, replicability and proof of possession of a specific body of knowledge. Even though HEs are much more time and resource-consuming, they are currently the accepted standard of reference for LLMs accuracy (33). Apart from this, no other agreements exist in the human evaluation of LLMs. Van der Lee et al. presented an overview of methods at the 12th International Conference on Natural Language Generation in Japan 2019. They approached the subject of sample size, demographics, design and number of questions, reiterating that concordance in the subject is scarce. In the following paragraphs, some of their conclusions will be presented.

First, there are two focuses on whether experts or non-experts will evaluate outputs. Evaluations could be more valid if judged by experts (34). Nevertheless, they could also have a bigger chance of introducing bias directly related to their fields of expertise. Additional bias is also described to be introduced by the exhaustion of the human evaluator, which could directly affect repeated questions if they are sorted statically, e.g., a set of questionnaires with multiple answers to assess in the same order could affect the perception and eval of the last questions. It is also difficult to have specialists agree or commit to revising a series of repeated questions, partly because of their clinical duties and the low amount of time they dedicate to other pursuits. In the end, both focuses are described as complementary (35), and the use of quantitative, statistical measurements of interobserver agreement is recommended (Cohen's kappa or Fleiss' Kappa), albeit measurements do not usually pass the 0.5 agreement (33).

Likert 5-point scales are the most widely used rating method in HE (36). Notwithstanding, scales that assign specific scores to answers and allow them to be ranked have shown better precision and are often more informative than the limited options of Likert. Presenting percentage agreements, though limited to descriptive approaches, and any free text evaluation resulting from the assessment is encouraged.

Human evaluations often include several evaluation frameworks: humans evaluating AI, humans evaluating humans, and AI evaluating AI. The latter is a response to the fact that HE tends to involve small numbers of experts since evaluating QA pairs in established datasets is numerous and virtually impossible to assess manually. In 2019, a paper by Marc Brysbae from Ghent University in Belgium stated that studies with sample sizes smaller than 50 individuals could be underpowered, and 100 or higher was a more ideal number of participants (37). This number of observations may be reached in smaller sample sizes by giving a single individual more than 2 questions to evaluate.

Evaluations are usually made on the content and structure of the output, which is checked for fluency, naturalness, quality, meaning preservations, relevance, grammatically, readability,

clarity, manipulation check, informativeness, correctness, syntaxis, qualitative analysis, appropriateness, not redundancy (33). Even so, there is a lack of structured and validated qualitative questionnaires for assessing these qualities, and they are developed according to each investigator's criteria. To address this issue, automatic metrics like faithfulness, coherence, semantic score and relevance were created and are continuously evolving. These aim to consider how a model constructs its responses from a given input and provide a more nuanced approach to how well models answer questions (38–40). However, direct human evaluation is still the standard of reference. Singhal et al. designed a three-point scale-based evaluation that included questions on agreement with scientific and clinical consensus, likelihood of possible extent of harm, reading comprehension, recall of relevant clinical knowledge, manipulation of knowledge via valid reasoning, completeness of responses, potential for bias, relevance and helpfulness. In this evaluation, 140 questions were randomly selected for assessment by 9 clinicians, and their outputs were bootstrapped to estimate significant variations. They evaluated answers from the Flan-PaLM and Med-PaLM models and clinicians. The results for each stance on each of the two models and for clinicians were as follows: 61.9%, 92.6% and 92.9% of answers were aligned with scientific consensus; 16.1%, 18.7% and 1.4% include content of great clinical significance; 47.6%, 15.3% and 11.1% have missing content of significant clinical relevance and 29.7%, 5.9% and 5.7% could potentially lead to harm respectively (31).

Mukherjee et al. carried out a large-scale study with an evaluation focused on the overall subjective user experience and targeted assessment of specific tasks or functions relevant to the CA's intended use. The subjective experience is considered by comparing human-to-human vs human-to-AI evaluation. They recruited 1100 US-licensed Registered Nurses and over 130 US-licensed physicians to evaluate a conversational agent. In this case, the agent would act as a healthcare provider and could call participants who would portray patients. Sixty randomly selected nurses would also interact with real patients, constituting the counterpart for comparison. The outcomes assessed were bedside manner, conversation quality, clinical readiness, the ability for patient education and potential for harm. Most questions were yes/no, while one question assessed user satisfaction and asked a blinded evaluator to rate the call on a scale of 1 to 10. The conversational tasks were evaluated with automatic methods or reviewed by nurses and clinicians according to their complexity. Some evaluated tasks included identifying prescriptions, correct dosage, contraindications for a particular drug, interpretation of laboratory values and vital signs, ability to recognize misspelt medical terms or medications and adequate answers to hospital policy questions (41).

3. EVALUATING EFFICACY AND HUMAN INTERACTIONS - CURRENT PERSPECTIVES IN LLM EVALUATION

Considering that there is still no generalized, validated strategy to test conversational agents' validity, the evaluation of LLMs and CAs has at least a more defined focus in terms of assessment, and that is on their ability to answer and reason correctly. Still, little is done to test their real-world efficacy and clinical impact. For CAs to succeed in health care, understanding effectiveness and task completion isn't enough. A review of 31 studies evaluating the efficacy and usability of CAs in healthcare found that few studies discussed real improvements to health care, cost-effectiveness or privacy and security outcomes (42). Future evaluation designs must prioritize this dimension of evaluation to assess the value generated by these tools correctly.

Additionally, user perception and real-world use are highly relevant as adopting new technologies depends on user willingness and satisfaction. Studies have shown that there is a "sociotechnical gap" described as "the discrepancy between what safety assessments predict in controlled, model-only settings and how models perform in the environments in which they are deployed" (43). In response, human interaction evaluations (HIEs) are emerging, focusing on assessing human-model interactions.

A study by Ibrahim et al. proposed a framework for designing HIEs specifically for safety assessment. The proposed evaluation focuses on the mechanisms of interaction, the outcomes or a combination of the two. This framework outlines the evaluation design in three steps: 1. Risk or harm identification, 2. Characterizing the use context, 3. Choosing evaluation parameters that can be applied to any risks identified for each use case. Some examples include assessing misuse, unintended personal harm and unintended external harm. It also outlines six human-LLM interaction models, namely collaboration, direction, assistance, cooperation, exposure and exploration, to guide evaluation accordingly (43). This approach has a high potential for thorough risk assessments but is costly compared to other evaluation methods.

Following the same avenue, Shuai Ma et al. have proposed an LLM-powered AI for AI-assisted decision-making, a framework to promote human reflection and discussion. It encourages humans to interact with the AI instead of simply accepting or rejecting the AI's suggestions (44). This shows a new approach to model assessment and underscores a needed focus on how humans and AI interact to make decisions, which is particularly relevant in healthcare. Their study highlights that this is particularly useful in challenging task cases for humans and AI. The framework comprises four essential activities: elicitation of thoughts, which encourages users to examine their reasonings and clarify their ideas and assesses the AI's ability to explain its "thoughts"; alignment of human-AI thoughts, which seeks to establish a common language for the two parties and to determine the extent of difference of opinion; discussion, were humans are encouraged to discuss with the AI to substantiate their views and clarify sources of evidence and the basis for a given suggestion and update of thoughts and discussion may expose reasoning flaws or conflicts, and it seeks to provide an opportunity for gaps to be considered and thoughts to be revised accordingly.

CONCLUSION

LLMs and CAs are becoming increasingly popular in healthcare, and the need for an appropriate strategy to validate them has become apparent. QA datasets based on medical licensing exams or other online sources are currently the benchmark for evaluating medical LLMs and CAs. While they are a good starting point for evaluating clinical knowledge and reasoning, they might be insufficient for risk and user assessment in real-life scenarios. In this sense, human evaluation is the preferred standard. However, these evaluations are costly and time-consuming, resulting in small sample sizes and a limited number of queries to evaluate. Current evidence suggests the best approach to assessment is the combination of human (qualitative) and automated QAs (quantitative) evaluations of AI models. Algorithms where other AI models are the leading evaluators, are being proposed in response to this known bottleneck. In this context, evaluating LLM and CA efficacy and human interactions also needs to be appraised in more realistic scenarios that permit the evaluation of the AI and its interaction with humans.

Contributions

NCV: Editor, Research and manuscript writing
IL: Research and manuscript writing
JM: Expert revisor
DJ: Expert revisor
MCV: Expert revisor
JZ: Expert revisor
LV: Expert revisor

Conflict of Interest

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organization for the submitted work; all authors are employed at Arkangel AI; no other relationships or activities that could appear to have influenced the submitted work.

REFERENCES

1. Barman B. The Linguistic Philosophy of Noam Chomsky. Philos Prog. 2014 Jul 15;51.
2. Fischer SR. History of Language. Reaktion Books; 1999. 244 p.
3. Larrañaga E. Hieroglyphics to Machine Learning: A Journey through Language Evolution [Internet]. Medium. 2023 [cited 2024 Aug 19]. Available from: https://medium.com/@EduardoLarranaga/hieroglyphics-to-machine-learning-a-journey-through-language-evolution-f5bbbbdfb8b0
4. The Rosetta Stone: Unlocking the Ancient Egyptian Language [Internet]. ARCE. [cited 2024 Aug 19]. Available from: https://arce.org/resource/rosetta-stone-unlocking-ancient-egyptian-language/
5. Muggleton S. Alan Turing and the development of Artificial Intelligence. AI Commun. 2014;27(1):3–10.
6. Furtado EL. ARTIFICIAL INTELLIGENCE: AN ANALYSIS OF ALAN TURING'S ROLE IN THE CONCEPTION AND DEVELOPMENT OF INTELLIGENT MACHINERY. 2018;
7. Chicco D, Jurman G. The ABC recommendations for validation of supervised machine learning results in biomedical sciences. Front Big Data. 2022 Sep 27;5:979465.
8. Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions | SN Computer Science [Internet]. 2021 [cited 2024 Aug 19]. Available from: https://link.springer.com/article/10.1007/s42979-021-00815-1
9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–40.
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The Rise and Potential of Large Language Model Based Agents: A Survey [Internet]. arXiv; 2023 [cited 2024 Jul 25]. Available from: http://arxiv.org/abs/2309.07864
11. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review | Journal of the American Medical Informatics Association | Oxford Academic [Internet]. [cited 2024 Jul 23]. Available from: https://academic.oup.com/jamia/article/31/3/746/7467291
12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 1877–901. (NIPS '20).
13. Feng K, Ding K, Ma K, Wang Z, Zhang Q, Chen H. Sample-Efficient Human Evaluation of Large Language Models via Maximum Discrepancy Competition [Internet]. arXiv; 2024 [cited 2024 Aug 20]. Available from: http://arxiv.org/abs/2404.08008
14. Bommasani R, Liang P, Lee T. Holistic Evaluation of Language Models. Ann N Y Acad Sci. 2023;1525(1):140–6.
15. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams [Internet]. arXiv; 2020 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2009.13081
16. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Trans Intell Syst Technol. 2024 Mar 29;15(3):39:1-39:45.
17. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics [Internet]. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002 [cited 2024 Aug 12]. p.

311–8. Available from: https://aclanthology.org/P02-1040
18. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out [Internet]. Barcelona, Spain: Association for Computational Linguistics; 2004 [cited 2024 Aug 12]. p. 74–81. Available from: https://aclanthology.org/W04-1013
19. BLEU is Not Suitable for the Evaluation of Text Simplification - ACL Anthology [Internet]. [cited 2024 Aug 12]. Available from: https://aclanthology.org/D18-1081/
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [cited 2024 Aug 13]. p. 4171–86. Available from: https://aclanthology.org/N19-1423
21. BioBERT: a pre-trained biomedical language representation model for biomedical text mining | Bioinformatics | Oxford Academic [Internet]. [cited 2024 Aug 13]. Available from: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506
22. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine [Internet]. arXiv; 2023 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2311.16452
23. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models [Internet]. arXiv; 2023 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2305.09617
24. Li J, Wang S, Zhang M, Li W, Lai Y, Kang X, et al. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents [Internet]. arXiv; 2024 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2405.02957
25. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns [Internet]. 2024 Mar 8 [cited 2024 Jul 15];5(3). Available from: https://www.cell.com/patterns/abstract/S2666-3899(24)00042-4
26. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In: Proceedings of the Conference on Health, Inference, and Learning [Internet]. PMLR; 2022 [cited 2024 Aug 13]. p. 248–60. Available from: https://proceedings.mlr.press/v174/pal22a.html
27. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering [Internet]. arXiv; 2019 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/1909.06146
28. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding [Internet]. arXiv; 2021 [cited 2024 Aug 15]. Available from: http://arxiv.org/abs/2009.03300
29. Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, et al. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain [Internet]. arXiv; 2023 [cited 2024 Jul 28]. Available from: http://arxiv.org/abs/2304.04280
30. Li J, Wang X, Wu X, Zhang Z, Xu X, Fu J, et al. Huatuo-26M, a Large-scale Chinese Medical QA Dataset [Internet]. arXiv; 2023 [cited 2024 Jul 28]. Available from: http://arxiv.org/abs/2305.01526
31. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172–80.

32. Reichenpfader D, Rösslhuemer P, Denecke K. Large Language Model-Based Evaluation of Medical Question Answering Systems: Algorithm Development and Case Study. Stud Health Technol Inform. 2024 Apr 26;313:22–7.
33. van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. In: van Deemter K, Lin C, Takamura H, editors. Proceedings of the 12th International Conference on Natural Language Generation [Internet]. Tokyo, Japan: Association for Computational Linguistics; 2019 [cited 2024 Aug 10]. p. 355–68. Available from: https://aclanthology.org/W19-8643
34. Amidei J, Piwek P, Willis A. Rethinking the Agreement in Human Evaluation Tasks. In: Bender EM, Derczynski L, Isabelle P, editors. Proceedings of the 27th International Conference on Computational Linguistics [Internet]. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018 [cited 2024 Aug 13]. p. 3318–29. Available from: https://aclanthology.org/C18-1281
35. Lentz L, Jong M de. The evaluation of text quality: expert-focused and reader-focused methods compared. IEEE Trans Prof Commun. 1997;40(3):224–34.
36. The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations | Request PDF [Internet]. [cited 2024 Aug 13]. Available from: https://www.researchgate.net/publication/338053074_The_use_of_rating_and_Likert_scales_in_Natural_Language_Generation_human_evaluation_tasks_A_review_and_some_recommendations
37. Brysbaert M. How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. J Cogn. 2019 Jul 19;2(1):16.
38. Rahimi H, Hoover JL, Mimno D, Naacke H, Constantin C, Amann B. Contextualized Topic Coherence Metrics [Internet]. arXiv; 2023 [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2305.14587
39. Parcalabescu L, Frank A. On Measuring Faithfulness or Self-consistency of Natural Language Explanations [Internet]. arXiv; 2024 [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2311.07466
40. Abeysinghe B, Circi R. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches [Internet]. arXiv; 2024 [cited 2024 Aug 10]. Available from: http://arxiv.org/abs/2406.03339
41. Mukherjee S, Gamble P, Ausin MS, Kant N, Aggarwal K, Manjunath N, et al. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare [Internet]. arXiv; 2024 [cited 2024 May 24]. Available from: http://arxiv.org/abs/2403.13313
42. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. J Med Internet Res. 2020 Oct 22;22(10):e20346.
43. Ibrahim L, Huang S, Ahmad L, Anderljung M. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks [Internet]. arXiv; 2024 [cited 2024 Aug 9]. Available from: http://arxiv.org/abs/2405.10632
44. Ma S, Chen Q, Wang X, Zheng C, Peng Z, Yin M, et al. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making [Internet]. arXiv; 2024 [cited 2024 Aug 9]. Available from: http://arxiv.org/abs/2403.16812