Using Generative AI to Answer Clinical Questions

September 2024

Katherine Eisenberg, MD, PhD, FAAFP

Al Stevens, PhD

Alan Ehrlich, MD, FAAFP

Introducing Dyna AI

The EBSCO Clinical Decisions team has introduced an Artificial-Intelligence feature for use with our point-of-care clinical decision support solutions, DynaMedex® and Dynamic Health™, called Dyna AI. Designed for healthcare professionals to ask complex clinical questions and get answers in a single step, the answers are generated from our highly curated, authoritative clinical content using the linguistic capabilities of generative AI. In this paper, we describe why this direct question and answer process is more efficient than traditional search, along with how we ensure the answers are accurate, high quality and free from the misinformation common in many AI systems.

Traditional Search

Since the inception of electronic information systems, the standard way that users find information they need has been the search function: the user enters a word or phrase in a search box, the entry is matched against a database of content, a list of matching topics is presented to the user, and then the user scans the list and clicks on the one that looks most relevant. The user reads the topic to find the needed information and if the information is not there, is not relevant or is incomplete, the user goes back to the list and clicks on another topic and then repeats the process.

Through the years, these techniques have been refined and modified. Initially, only topics that matched the terms exactly were returned. Later, word-stem matching was incorporated so that terms like “surgery” would also match “surgeries”. Further refinements, called semantic search added the ability to match synonyms and abbreviations, so that “atrial fibrillation” would match “afib”. Most recently electronic information systems have incorporated knowledge graphs that use richer relationships between terms. This enables searches for “sertraline” to also return topics related to it, either more generally, like “ssri” or related in other ways such as tradenames like “Zoloft”.

Semantic search techniques made searching easier, returning more relevant topics but while they relieved the user of needing to guess the best terms, they did not change the fundamental search-and-sift paradigm. While these techniques work well for users looking for multiple sources to read and study, they do not work well for users who want to ask a question and get an answer. To get an answer to a specific question, users of traditional search are still left with a list of topics to select, open, and read.

Large Language Models and Generative AI

A medical information-system user is almost always looking for an answer to a question. Often these questions include context and detail. Examples range from relatively simple questions like:

> *What are the recommended antibiotics for treating community-acquired pneumonia (CAP)?*

To more complicated questions like:

> *What is the target HgbA1c level for a 78-year-old patient with type 2 diabetes mellitus and mild cognitive impairment (MCI)?*

To search and sift for answers to these questions in a database of topics is not easy and requires time and effort. For the first one, should the search start with CAP or pneumonia, or community-acquired pneumonia? Should the search be "target hemoglobin A1c in type 2 diabetes", "diabetes management in older patients" or something else? What if these searches give different results? The amount of time and effort to find the information desired may be unrealistic in a busy clinical setting, ironically, when the information is needed most.

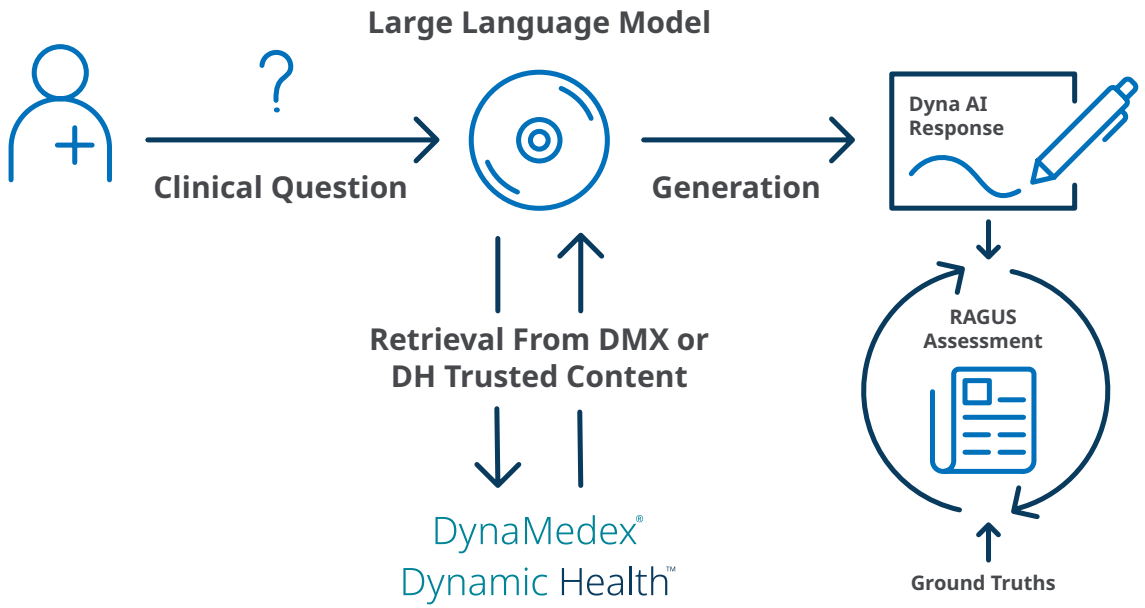
We saw the potential of generative AI to enable users to skip the search-and-sift process and just ask their question and get an answer. While asking a direct question is a new workflow for users, it is more in line with a typical case-based clinical thought process than the traditional search approach. We implemented this Q&A model in our Dyna AI system by using a large language model to respond to the user's question, retrieve the needed information from our curated content and then use it to answer the question directly. We go further by including the ability to interpret the user's question from shorter abbreviated phrases, including medical abbreviations like CAP, that reflect the way clinicians want to ask questions, and which can be easily typed into a search box. Dyna AI interprets the user's question from the entered query, shows the user the interpretation and then answers it, all in a matter of seconds and with a view towards the needs of healthcare professionals.

There has been a lot written about AI systems in the last year. Much of it raises concerns about trust and bias, which likely stems from an incomplete understanding of how AI systems work. Skilled, multidisciplinary development teams who have in-depth understanding of how these systems work, who can carefully evaluate them and implement the latest techniques to restrict the accessible information sources can deliver AI solutions that dramatically improve the time it takes clinicians to access critical medical information. At the same time, products designed

to improve AI literacy among clinical users can support clinicians with safe, responsible AI adoption. As users better understand how different AI systems work, they can better understand the risks and benefits of these systems and better position themselves to benefit from their use.

To assist in this understanding, we describe how Dyna AI works to accurately answer clinical questions using high-quality evidence. Using a Large Language Model (LLM) and Generative AI to interpret user queries, Dyna AI retrieves the relevant, trusted information that has been curated by our team of several hundred expert editors, and then summarizes that information to respond with a succinct answer in seconds. This way of using Generative AI with a Large Language Model is called Retrieval Augmented Generation, an industry standard way of providing answers based on a selected set of content.

Dyna AI | Retrieval Augmented Generation



An LLM contains a large array of numbers, called parameters – a very large array. Open AI's GPT4, which has dominated the news throughout 2024, uses almost 2 trillion parameters to represent the content it contains. If you filled a spreadsheet the size of a US football field with numbers, you would need more than 20,000 football field-sized spreadsheets to store those parameters.

When an LLM is initialized, the parameters are set to random values. The model is then "trained" by having it read text. As it reads the text, it predicts the next words based on the text it has already read. When it correctly predicts the next words, the parameters are adjusted in one direction. When the prediction is incorrect, the parameters are adjusted in another direction. As the parameters are adjusted from reading more and more text, the predictions get better and better and will then more closely match text the model has never read. Training a model with this many parameters requires a lot of text. For GPT4, if the training text were printed in books, it would fill a shelf several hundred miles long.

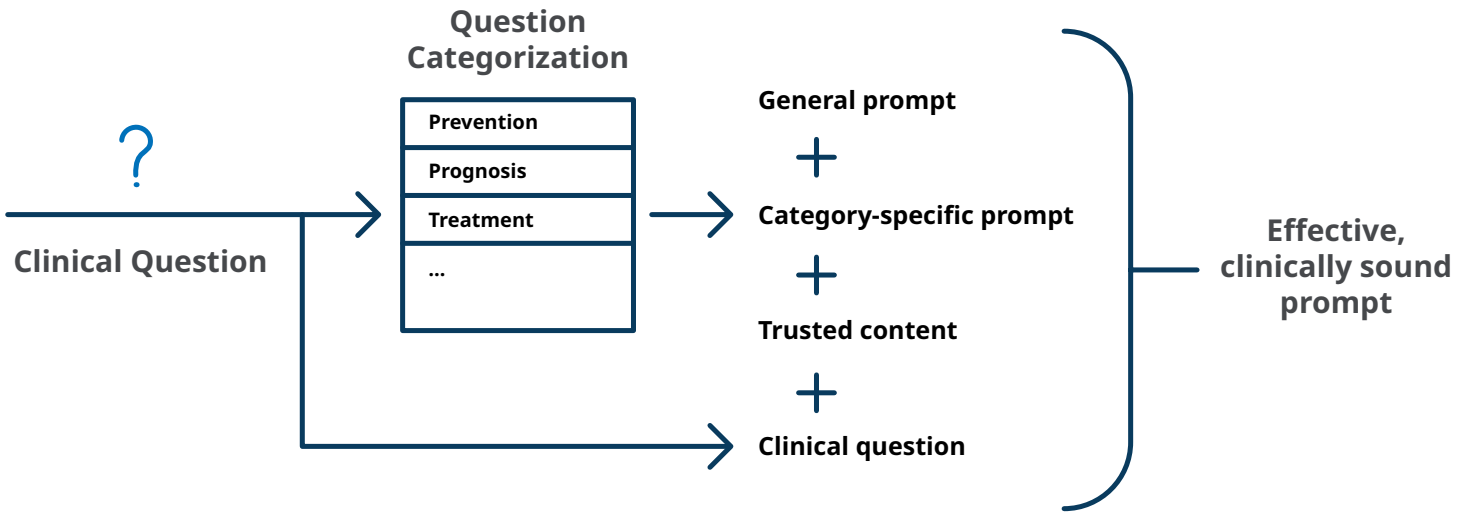
After training, the parameters of an LLM represent the information in the content used to train it. Since much of the training material comes from open Internet sources, it can include facts, opinions and falsehoods. The parameters also represent the structure and rules of language such as generation, summarization, and transformation.

Generative AI uses a trained LLM to generate new content. When used with text, a generative-AI system reads text it has never read before and based on the complex relationships within the LLM parameters, generates new text in the same way it predicted text during the training process. The previously unseen text is called a "prompt". A prompt can be a few words, or it can be many pages, or even an entire document. The newly generated text is called a "response". The response can be long, short, formatted in different ways, or even in different languages. The response format, like the response content, is determined by the prompt.

Using an LLM and Generative AI effectively requires using its rules of language to provide information, ideally from a restricted set of trusted content.

Dyna AI does this. It uses the linguistic information in the LLM to expand the words or phrases from the users' queries into complete clinical questions. The LLM categorizes the question into one of several clinical categories, developed by our clinical staff, and we then construct a prompt that includes that question along with the relevant content sourced directly from our own curated, trusted content. We use this prompt to direct the LLM to summarize the content using language meaningful to a clinician. By limiting the LLM to summarizing only from our trusted content, we avoid the issues seen when the LLM response also uses the content it was originally trained on, which may be wrong or misleading in a medical context.

Dyna AI | Constructing an Effective Prompt



In simple terms, Dyna AI uses the linguistic capabilities of an LLM to interpret the user's query as a question, find the relevant content, read it, summarize it and present it to the user.

The DynaMedex and Dynamic Health Content

Critically, Dyna AI answers depend on high-quality content. It's not sufficient to simply base answers on a selected set of journals. Even in the best journals, the content varies in quality and relevance to medical practice. Dyna AI answers come from the content in EBSCO's point-of-care clinical decision support solutions, DynaMedex and Dynamic Health. A team of several hundred editors use a rigorous multi-step process to systematically identify, select, evaluate, determine what is relevant, synthesize, derive and update conclusions from the published medical literature. The results of this process are published daily in DynaMedex and Dynamic Health. This highly curated content is used as the basis for Dyna AI.

An Example of Dyna AI in Action

In Dyna AI, users enter their query in a search box. For example:

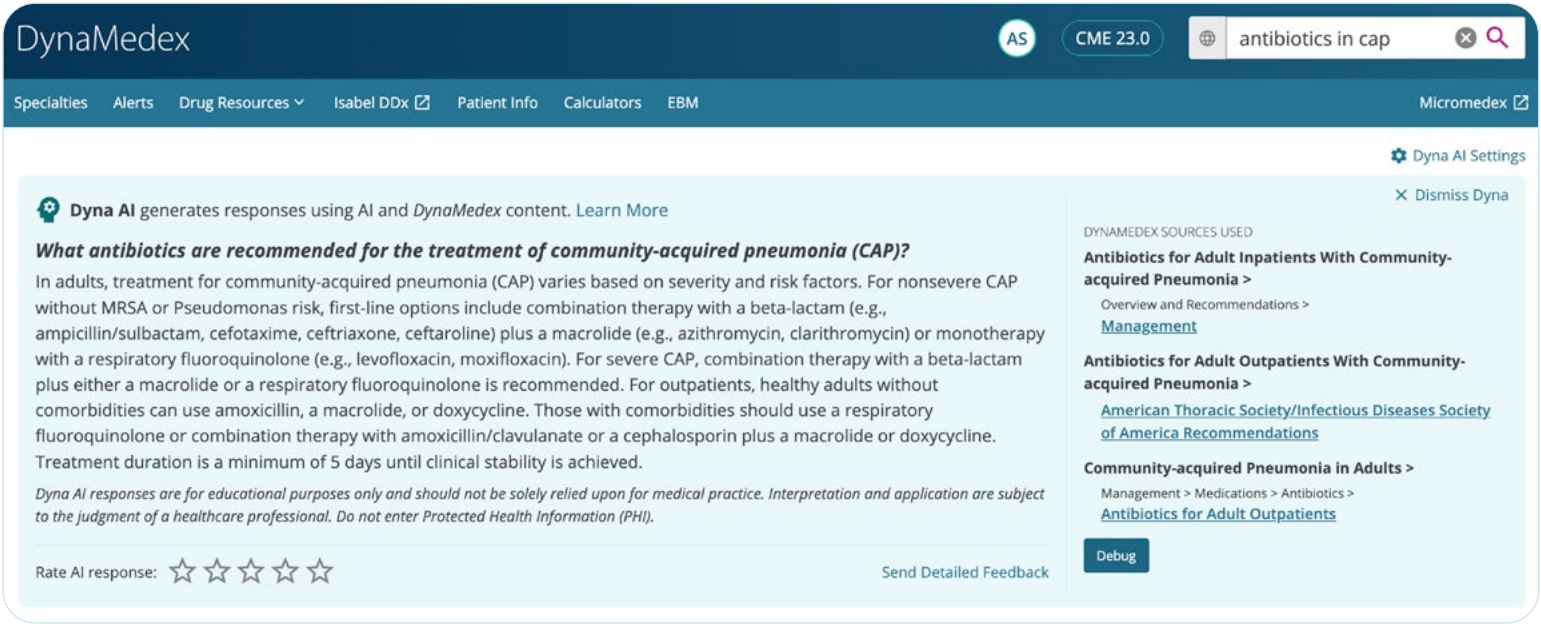
antibiotics in cap

Dyna AI combines this query with sample questions and then uses the LLM to expand this short query into a formal English question. This question is designed to minimize ambiguity and express the query with more clarity. We show this to the user to confirm the intent has been correctly interpreted. The expanded query is:

> *What antibiotics are recommended for the treatment of community-acquired pneumonia (CAP)?*

This question is then used to retrieve the relevant content from DynaMedex. To answer this question, four different sections from three different topics are retrieved. These sections, along with the question and additional linguistic guidance are assembled into a prompt and sent to the model. The Dyna AI response is:

> *In adults, treatment for community-acquired pneumonia (CAP) varies based on severity and risk factors. For nonsevere CAP without MRSA or Pseudomonas risk, first-line options include combination therapy with beta-lactam (e.g., ampicillin/sulbactam, cefotaxime, ceftriaxone, ceftaroline) plus a macrolide (e.g., azithromycin, clarithromycin) or monotherapy with a respiratory fluoroquinolone (e.g., levofloxacin, moxifloxacin). For severe CAP, combination therapy with a beta-lactam plus either a macrolide or respiratory fluoroquinolone is recommended. For outpatients, healthy adults without comorbidities can use amoxicillin, a macrolide, or doxycycline. Those with comorbidities should use a respiratory fluoroquinolone or combination therapy with amoxicillin/clavulanate or a cephalosporin plus a macrolide or doxycycline. Treatment duration is a minimum of 5 days until clinical stability is achieved.*



This response is a summary generated only from our trusted content by Dyna AI using the linguistic knowledge encoded in the LLM, guided by examples in a prompt that included the expanded query along with examples of appropriate responses.

Complex Queries

In many cases, formulating a search using a classic search system is itself a challenge. Consider this query:

> *What is the target HgbA1c level in a 78-year-old who has type 2 DM and MCI?*

The answer could be in a topic on Type 2 Diabetes, in a section on glycemic goals. It's unlikely to be in a section that includes the terms "78-year-old". Dyna AI can retrieve and summarize information like this without requiring the user to enter search terms that match the content.

Dyna AI first expands the query to include the spelled-out abbreviations:

> *What is the target HgbA1c level for a 78-year-old patient with type 2 diabetes mellitus and mild cognitive impairment (MCI)?*

It then retrieves content from three different topics and produces this response:

> *For a 78-year-old patient with type 2 diabetes mellitus and mild cognitive impairment (MCI), a less stringent HbA1c target of < 8% (64 mmol/mol) is recommended. This target is appropriate due to the patient's cognitive limitations and the potential risks associated with more aggressive glycemic control, such as hypoglycemia. The goal is to balance glycemic control while minimizing the risk of adverse effects from diabetes medications. Regular monitoring and individualized care are essential to avoid hyperglycemia leading to symptoms and acute complications.*

As in the prior example, Dyna AI first expands the query. In this case, it expands the acronyms as spelled out medical terms and clarifies that "78-year-old" refers to a patient and determines that patient is a member of a class, in this case patients over 65 years old. The retrieval and summarization process generates this answer.

Ensuring Clinical Quality

Generating answers to clinical questions introduces an additional layer of content transformation that can, if not carefully done, result in incorrect or even dangerous answers.

Dyna AI bases its answers on our trusted, curated content using industry standard techniques to mitigate this issue, but there are additional considerations that need to be managed to ensure high clinical quality. Generative AI using large language models is powerful, but it is still just a tool that needs to be both designed and used carefully and correctly.

LLM model changes, modifying the prompts, adding content, removing content, and adjusting the retrieval process all affect the quality of the answers. To achieve and maintain clinical quality requires that clinicians work closely with technicians and that appropriate review and advisory boards are in place and integrated with the development process.

To ensure the highest quality possible, our development teams include clinicians working directly with developers. The clinicians on the team manage a group of additional specialists reflecting diverse medical perspectives to continuously evaluate answers, focusing especially on features that were changed, but also ensuring that this has not resulted in loss of previous quality. The metrics used continue to evolve but include several dimensions of clinical quality such as accuracy, relevance, completeness, tone, and inclusion.

The deep involvement of our clinical team in every element of the Dyna AI build, design, and clinical quality assurance process reflects our commitment to building an AI-powered system that is safe, usable, and supports clinicians with access to the highest-quality information. Because of the nature of generative AI outputs in clinical medicine, this approach requires many clinical experts to be involved with providing feedback. Given that evaluation of generative AI applications in healthcare is an evolving area, we have contributed our experience in this area to multiple working groups of the Coalition for Health AI.

As one of our clinical safety measures, we have established that not all questions can be answered using Generative AI. Based on input from our clinical team, we use a set of clinically reviewed guardrails to intercept questions that would generate answers that do not pass our quality metrics. When these questions are encountered, we log them for further analysis and provide the user with an explanation as to why the answer is not available. As the technology and system advance, our guardrails evolve as well.

The deep involvement of our clinical team in developing Dyna AI has also allowed us to design the system with clinical workflows and needs in mind. We know that decision support tools like Dyna AI must augment and support clinicians' needs and judgment by improving the speed at which users can access high-quality, evidence-based clinical information.

Clinical Quality Monitoring – Continuous Improvement

To integrate and continuously improve clinical quality, we use a set of techniques called the "retrieval augmented generation assessment suite," usually referred to by the acronym RAGAS. As our clinicians evaluate responses, we collect the ones rated at the highest level. We have the clinicians correct those with lower ratings. We call this set of high-quality answers our golden set. We then have Dyna AI generate answers to the golden-set questions and make further use of the LLM linguistic capabilities to compare the results with the golden-set answers. This comparison provides several metrics that enable us to instantly see how a change affects clinical quality and if it does, pinpoint the reason. As the golden sets grow, we can continuously improve the overall quality and expand the breadth of questions that Dyna AI can answer.

In addition, our robust clinical evaluation to stand up Dyna AI has allowed us to build the infrastructure and expertise to transition into continuous clinical quality monitoring. Our golden sets, database of responses with clinical evaluator comments, information flows to the developer teams, and reporting tools provide an opportunity for real-time monitoring of the Dyna AI system. It is easy to assume that a technological advance like generative AI can be fully automated, but our stance is that for a healthcare application clinicians need to be directly involved in the clinical quality process on an ongoing basis.

The Future

Dyna AI in its current form directly answers clinical questions from trusted content. Looking only a little way into the future, we can see that closer integration with the clinical workflow, providing additional context from the clinical setting of the user, and further personalizing the responses will enable us to provide even more useful answers. LLMs were originally developed to translate from one language to another. These linguistic capabilities provide the opportunity to answer questions in languages other than the language of the source content. Next generation LLMs will include the ability to integrate calculators and algorithms providing opportunities to ask and get answers to even more complex questions.

Learn More about GenAI

For those who want to dive deeper into LLMs, Generative AI and RAG, we suggest these three papers.

How Large Language Models Work, Andreas Stoffelbauer
Describes how LLMs work in some detail, striking a balance between superficial and deep technical descriptions.
https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

What is ChatGPT Doing … and Why Does it Work, Stephen Wolfram
A comprehensive description of how ChatGPT works. It gets more technical as you read on, but you can stop when you want and still come away with a good understanding.
https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work

Augmented Generation for Knowledge-Intensive NLP Tasks, Patrick Lewis et. al.
The paper that introduced retrieval augmented generation. It's very technical, explaining in detail how the RAG retrieves content from a controlled set and generates answers to questions.
https://arxiv.org/pdf/2005.11401