Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: November 30, 2023

Citation: Macfarlane D., 2023. Professional Report Generation Using Lexeme Theories Versus Openai's Generative Pretrained Transformer, GPT-4: A Comparison, Medical Research Archives, [online] 11(11). https://doi.org/10.18103/mra.v11i11.4700

DOI
https://doi.org/10.18103/mra.v11i11.4700

ISSN: 2375-1924

REVIEW ARTICLE

Professional Report Generation Using Lexeme Theories and Openai's Generative Pretrained Transformer, GPT-4: A Comparison

Donald Macfarlane, MD, PhD,

University of Iowa and Lexeme Technologies, LLC

Email: donald-macfarlane@lexetech.com

ABSTRACT

The public release of OpenIA's GPT-4 has caused an explosion of interest in the ability of large language models to generate text documents in response to a simple text prompt. These documents can appear to be genuine professional reports, such as medical case notes. Expert-written templates guided by lexeme theories (TGLT) is a system under development which creates professional notes exploiting a set of theories and a lexicon which converts a clinician's ideas into text. We explored the differences between the two systems to determine if they can be used in clinical practice. Every element in a document created by TGLT is triggered by the user, whereas GPT-4 created documents may include invented text. LGLT can generate complex clinical notes that are more complete, more orderly, and less error-prone than conventionally written notes. The lexicon constructed for TGLT can be updated or corrected rapidly by end-users, whereas GPT-4 uses a huge library that may take many months to update. TGLT notes are concise, complete, and organized in a defined order, whereas GPT-4 may be incomplete and poorly ordered. TGLT can alert the user to recent best practice advisories. TGLT issues computer codes for every text element in the document and does not include confidential identity information, enabling the facile aggregation of the content of notes for down-stream analysis. GPT-4 does not issue computer codes, and generates text that may include patient identifiers. TGLT costs vastly less than GPT-4. We conclude that TGLT has none of the manifold disadvantages that GPT-4 has for creating professional reports.

Introduction

The recent release of Open AI's generative pretrained transformer GPT-4[1], using a large language model, has generated an intense examination of its remarkable ability to create highly plausible documents in response to simple prompts. GPT-4 would appear to be extremely valuable for public-facing activities, such as communication with customers and creative writing, but the possibility that a document created by GPT-4 could be submitted as a professional medical note creates immediate concerns.[2],[3] Should such a document be treated as a valid source for medical billing, and can anyone rely on its veracity?

Clinicians write about 1 billion medical reports concerning outpatient visits per year in the United States,[4] and likely a similar number of inpatient reports. These notes provide a record of an interaction with a patient, generally including the history and physical findings of the patient and the clinician's opinions concerning the patient's management. Notes often include information from laboratory or imaging studies and consultants' reports. The notes inform future clinicians caring for patients. They also provide a basis for billing for medical services, and provide key evidence in the defense against an allegation of malpractice. They may also be used for clinical research and administrative decisions.

For centuries, medical notes were handwritten by the clinician and aggregated into a binder. Medical notes today are stored in an electronic medical record system, but the body of the note is still written by the clinician often with the aid of a word processor, a speech-to-text generator, or a human scribe.

Notes created today are much bulkier than they were,[5] and they continue to contain errors of fact, omission, reduplication, grammar, and spelling. They are often poorly organized. These error-filled notes cannot be easily analyzed by computer: many decades of effort to establish natural language processing for this purpose have been largely unsuccessful.[6] As such, the typical medical note today cannot be effectively analyzed by computer algorithm, necessitating expensive and time-consuming reading of the note by a human expert to determine its content.

The purpose of this article is to compare the validity of document preparation using TGLT with that of the generative pretrained transformers, GPT-4. In particular, we evaluated the products to assist in the determination as to whether they can be used in clinical practice.

Methods

We evaluated medical note generation by Open AI's GPT using a commercially available interface operating via an application program interface (*ChatOn AI*, from AIBY). We evaluated note generation by TGLT as performed in the in the clinical trial.[7]

Expert-written templates guided by lexeme theories (TGLT)

To find a solution to computerizing the entire content of medical notes, we developed a system which reverses the thinking that is the basis for natural language processing. Instead of trying to extract the ideas from clinician-created text, TGLT solicits ideas from the clinician and creates text from expert-created templates selected using a set of linguistic theories about how we write professional reports. It uses a special lexicon of templates that includes text fragments which can be included in the final document. The resulting computer system enables its user to write medical notes ready for printing or uploading to an electronic medical record. TGLT successfully generated notes of great complexity within a pilot clinical trial. These notes were more complete, more accurate and less error-prone than conventionally written notes, and they were created at about the same speed as dictation. Every element in the note was identified with a computer code, and each was demonstrably triggered by an action of the user. An example of an LGTL note can be viewed here.[7]

Lexeme theories

We developed the lexeme theories starting with a simple concept -language consists of a stream of small and discrete units of information. We use the term "lexeme" to denote this atomic unit. We found it valuable to divorce the informational content a lexeme from the text that is used to convey it. This frees us to decide how we want to express the information linguistically, for example in an abbreviated or expanded text style and which language to use. This informational unit is identified by a unique computer code. We further posit that each lexeme can be split into two components: the subject matter or topic, which can be expressed as a question (or "lexeme query"), and one of a set of answers to that query (or "lexeme responses").

The lexeme theories hold that 1) all the lexemes queries required to write a professional report can be placed in one acceptable order ("coherence"), 2) that the need within the report for a particular lexeme query is restricted by the context within which the report is written and by the responses already selected ("predicance"), and 3) by the

level of detail needed ("level"). These theories power the logic engine of TGLT.[8]

TGLT works iteratively in coherence order, presenting a starter lexeme query associated with the user's login information with its set of responses. When the user selects a response, 1) the text fragment associated with that response is added to the output document, 2) any predicants associated with the response are added to the system's predicants list, and 3) the response's level is set. The system then searches in coherence order through the lexicon to find the next lexeme query with a predicant match and an adequate level and presents that query to the user with a collection of responses. This systematic and orderly approach ensures that a proper path is taken by the user to complete a note.

The current iteration of TGLT runs on a browser on the user's device. It downloads several blocks of the lexicon in advance of need, and it does not need a consistent connection to the internet. The lexicon is written by users who have been trained to be authors, lightly overseen by an editorial office.

Artificial Intelligence and Generative Pre-trained Transformers

Artificial intelligence is widely used to predict the next word in a text stream, greatly improving the accuracy of speech-to-text and typographical error correction. Advances in computer design have using large language models led to the ability not just to predict the next word, but to create an entire document from a simple text prompt. These technical advances include the concept of a transformer, a software module that can examine huge volumes of text non-recursively using multiple parallel attention systems and a time-limited memory to generate parameters. These parameters can then be used by the same transformer to guide the generation of text.

Open AI released GPT-4 on March 14, 2023.[1] It was pre-trained by using a supercomputer with 285,000 cpu's to examine a huge corpus of written information from online and other sources. [9] The size of the resulting library enabling its operations is not published, but the preceding GPT-3 houses 175 billion parameters. GPT-4 uses context windows for prompts accommodating up to 32768 tokens, the equivalent of about 50 pages of text. The system is reinforced by human feedback, reducing the risk of generating harmful text.[10]

The purpose of this article is to compare the process of document creation by TGLT and GPT-4.

Veracity of text

Open AI's GPT-4 was not created to prepare professional reports, but it creates documents that can certainly masquerade as genuine. For instance, when we entered the prompt "*Prepare a clinical note of a first visit by a 70y male who needs left knee replacement*", GPT-4 generated the following document: image.png (672×738) (helprace.com)

Note that this document could certainly masquerade as a genuine clinical note and would likely be accepted if submitted in support of billing. Note also that it contains about 50 assertions of fact - of which about 47 are completely fabricated. This document can be compared to the document prepared by LTGL concerning a patient with hemophilia, shown here[11].

Size of Library

As indicated above, the size of the library needed by GPT-4 is huge. In contrast, our rough estimate is that the entirety of medical practice can be provided by a few million lexeme queries of about 2 kilobytes each.

Quality of library

GPT-4 was trained using a vast compendium of the world's literature. Much of this literature was not generated with the sole intention of revealing the truth: it was written to maximize profit, to advance a political point of view, or innocently to recount a conventional but misguided understanding. The recent alliance of Microsoft and Epic[12] may lead to pretraining using a large collection of existing patient notes. This will enable data mining which will yield very valuable insights into current practice across medicine, (including occasionally what therapy yields the best results). But (unless guided by clinical experts) this data collection will be quite out of date and will include all the common errors in medical practice.

Timeliness of library

TGLT accesses a lightly curated lexicon which can be updated very rapidly. For instance, lexeme queries addressing what drugs to use to treat a certain disease can include information about a new drug (including its indications, efficacy, dosing, side effects, drug interactions, and recommended laboratory monitoring) or new recommendations about disease management can be written and uploaded to the lexicon in an hour or two. This new information will immediately be offered to all the relevant users as responses that can be selected. Thus, TGLT can be updated as fast as writing a press release.

In contrast, the GPT-4 library will need to be completely retrained to assimilate new information. This is a computationally intensive process, suggesting a time lag of months to several years between retrainings.

Error Correction

Within TGLT, untrue statements can only arise by a user making an incorrect selection, or by an error in the lexicon. A user who finds an error in the lexicon can simply overwrite the incorrect information in the output stream of the record being generated, and that overwrite and its context is sent to the editorial office to initiate the correction of the lexicon. A user who is authorized to be an author can change the lexicon more extensively. The lexicon used by TGLT can be updated many times a day, and these updates will permeate to all relevant users rapidly.

In contrast, if GTP-4 generates text with an untrue statement, there is no obvious way that a user can correct the information to prevent propagation of the error because the pre-training data used by CPT-4 is vast, hugely expensive and static.

Completeness and Orderliness

The user of TGLT is required to work through the document in coherence order, ensuring that the user will be presented with all the lexeme queries that the authors of the lexicon think necessary. As a result, the notes are complete and will follow a prescribed order, making the notes much easier to read. The orderliness of the process prompts users to address issues that might otherwise be overlooked, and lexeme responses can include best practice advisories.

GPT-4 has no assurance of completeness or orderliness.

Accuracy

Given the entry of the same responses, TGLT generates one clinical note: it makes no attempt to produce a varied result. This note will be true if the user makes the right selections and if the lexicon is written correctly.

On the other hand, GPT-4 uses a probabilistic approach to understanding the prompt it is presented with. This alone produces variability in the text it generates. GPT-4 and GPT-3 freely embellish their output with invented facts to enhance the narrative value of the product. This is a nice touch for creative writing, but such "hallucinations" have no place in a medical note.

Ability to enter subjects into clinical trials

Users of TGLT are guided by the organization of the lexicon. They can only reach a lexeme query if certain conditions are met. For instance, a pediatrician seeing a patient with symptoms suggestive of strep throat will encounter a query addressing treatment. The lexicon could include an option to print an informed consent document and randomize the patient between therapeutic options, enabling very facile double-blind randomized controlled trials.

Computerized result

Every lexeme in a note generated by TGLT is identified by a computer code. These codes may have quite subtle meanings, and some codes may be replaced by better lexemes as the lexicon evolves. A collection of TGLT expressed in codes enables a very facile search for correlations. For instance, the notes can be divided into two categories (such as notes concerning patients who improved, and those who deteriorated), and the lexeme responses that correlate with this outcome can be identified.

Such facile data analysis will not be possible with GPT-4, since it produces elegant text but no codes for scrutiny.

Cost to develop

Open AI received over $1 billion in investments to create GPT-4 and has future pledges of an additional $10 billion. In contrast, the cost of developing TGLT and writing a lexicon sufficient for the pilot trial is in the low millions. Lexeme queries cost about $50 each to write. A lexicon targeted to a specific area of practice may require a few thousand lexemes at a cost of much less than $1 million. An essentially complete lexicon to cover the entirety of medicine may cost less than $50 million to create.

Confidentiality

TGLT does not include any of the 18 identifiers that are protected by HIPAA in its output.[13] A note generated by TGLT is intended to be ported to the user's EMR, where any necessary text identifying the patient and other personal information can be added semi-automatically within the EMR. As a result, a collection of TGLT expressed as computer codes does not need to be de-identified before analysis by a third party.

If GTP-4 is loaded with a prior clinical note, then the inclusion of HIPAA identifiers is likely.

Conclusions

We conclude that medical notes written by GPT-4 include fabrications that cannot be accepted as the documentation of a clinical interaction intended to memorialize the event, to support billing or in the defense of malpractice.

In contrast, professional notes created by TGLT are accurate, up to date, complete, and informed by best practice. TGLT generates computer codes identifying every item in the resulting note, and they do not include patient identifiers, enabling facile data aggregation. Every item in the document is created as a result of a choice made by the user, so the note can be submitted as a valid note for billing purposes.

Conflict of Interest Dr Macfarlane owns intellectual property underlying LGLT.

References

1. OpenAI. GPT-4 Technical Report. *arXiv*:2303: 08774, 2023
2. Lee P, Bubeck S, et al. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med* 2023; 388:1233-1239. https://www.nejm.org/doi/full/10.1056/NEJMsr2214184
3. Andrea Fox. GPT-4 and LLMs like it hold promise for healthcare, but caution is warranted. *Healthcare* IT News April 14, 2023. https://www.healthcareitnews.com/news/gpt-4-and-llms-it-hold-promise-healthcare-caution-warranted.
4. Santo L, Kang K. National Ambulatory Medical Care Survey: *National Summary Tables*. https://dx.doi.org/10.15620/cdc:123251. 2019
5. Liu, J *et al*. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J Biomed Info,* 133:104149, 2022
6. Vidiswaran, VGV, *et al*. Special issue of BMC medical informatics and decision making on health natural language processing. *BMC Med Inform Decis Mak*. 19(Suppl 3):76. 2019
7. Pilot trial of semi-automated medical note writing using lexeme hypotheses. Gugel D, Lentz S, Perepu U, Sharathkumar A, Staber J, Sutamtewagul G, Macfarlane D. *Int J Med Inform*. April 2020 136:104095
8. The lexeme hypotheses: Their use to generate highly grammatical and completely computerized medical records. Macfarlane D. *Med Hypotheses*. 2016 92:75-9
9. Bushwick, S. What the New GPT-4 AI Can Do. *Scientific American*, March 16, 2023. https://www.scientificamerican.com/article/what-the-new-gpt-4-ai-can-do/
10. Vaswani, *et al*. Attention Is All You Need. *arXiv*:1706.03762
11. Pilot trial of semi-automated medical note writing using lexeme hypotheses. Gugel D, Lentz S, Perepu U, Sharathkumar A, Staber J, Sutamtewagul G, Macfarlane D. *Int J Med Inform*. April 2020 136:104095
12. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI *Microsoft News Service*. April 17, 2023. https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/
13. Art Gross. The 18 PHI (Protected Health Information) Identifiers. *HIPAA Secure Now*, May 16, 2022. https://www.hipaasecurenow.com/the-18-phi-protected-health-information-identifiers/

Lexeme based vs ChatGTP, revised.pdf 362.8 KB