Chat GPT 활용 연구의 계획과 출판까지의 과정: 프론티어로서의 여정

삼성서울병원 영상의학과 이정현

1. 서론

대형언어모델(large language model; LLM) 및 이로부터 파생된 대화형 인공지능 모델들은 OpenAI 社에서 2022 년 공개한 서비스 ChatGPT 를 기점으로 폭발적인 대중의 관심을 받기 시작했고, 이어서 2023 년 공개한 GPT-4 를 통해 영상의학을 포함한 다양한 분야에서 실제로 활용이 가능할 정도의 추론 능력까지 도달할 수 있음이 증명되었다.[1,2] 잘 알려져 있는 LLM 들은 대화에 특화된 모델들로, 언뜻 보면 의료에서의 활용이 제한적이라 생각할 수 있지만 LLM 에 대해 더 깊은 이해를 가진다면 영상의학을 비롯한 의료현장에 접목해 볼 수 있는 방법이 상당히 많다는 것을 알 수 있다.

2. 대형언어모델

LLM 이라는 용어는 널리 쓰이지만 정확한 정의는 어려우며 이르게는 2000 년대에도 문헌에 등장한다.[3] LLM 에 의미에는 모델의 구조에 관한 것은 포함하고 있지 않지만, 2017 년 발표된 transformer architecture 에 관한 논문이 현대적인 의미의 LLM 발달의 첫 이정표라고 할 수 있으며, 이 논문에서는 번역 작업에서 transformer 의 성능을 검증하였다.[4] 이 당시에는 “large”가 포함되지 않는 language model 의 개념이었고, 이후 매개변수의 개수를 늘리면서 작은 모델들에서 관찰되지 않은 많은 능력들이 드러나게 되면서 large language model 이라는 개념이 형성되었다. 그렇기에, 현재 통용되는 LLM 의 개념은 모델의 파라미터 개수가 많은 것 뿐만 아니라, 특정 작업에 대한 지도학습 (supervised learning) 없이 다양한 작업을 잘 수행할 수 있다는 개념도 포함한다고 볼 수 있다.[5] 이처럼 챗봇 (chatbot) 서비스를 위한 LLM 은 다양한 LLM 의 응용 방법 중 하나일 뿐이며, LLM 은 실제로는 언어, 엄밀히 말하면 텍스트로 표현되는 언어와 관련된 모든 작업에 적용할 수 있는 것이다. 의학 및 과학 분야에서 필수적인 논리적 추론도 언어로 표현할 수 있기에, GPT-4 와 같이 고도화된 LLM 은 인간과 유사한 수준의 논리적인 사고를 모사한다.[2] 한발 더 나아가 생각하면, 컴퓨터와 모바일 기기 등에 사용되는 코드도 언어의 일종으로, LLM 을 통해 소프트웨어 제어 또한 자연어로 가능하게 될 수 있다.[6]

3. 의료에서의 대형언어모델

a. Medical Question Answering and Comprehension

가장 활발하게 연구가 이뤄진 분야이며, 주로 LLM 의 의학 지식과 의학적인 추론 능력을 시험해보는 연구들이 이루어졌다.[7-9] 전반적인 의학 지식을 시험하기 위해 사용된 데이터는 각종 의사면허시험 모의 문제 및 언어모델들을 위해 특별히 제작된 문답 형식의 텍스트이며,의학 분야에 특화된 기존 작은 모델들보다 뛰어난 성능을 보이기도 했으나, 아직까지 범용 대형언어모델의 in-context learning (프롬프트를 통해 학습하는 능력)이 특화된 언어모델들보다 우수한 것이 맞는지에 대한 논란이 있다.[10] 저자의 경험으로는 단순한 의학 지식으로 답변 가능한 문제들은 특화된 작은 모델들의 성능이 좋을 수 있으나, 의학적인 추론이 바탕이 되어야하는 고도의 문답에는 대형언어모델에서만 나타나는 추론 능력이 필요하고, out-of-domain 데이터에서의 일반화가 가능하기에 대형언어모델의 장점이 더 크다고 생각한다.

이러한 형태의 활용은 환자 상담, 감별 진단 작성, 의학 지식 검색에 적용할 수 있으며 특히 환자 상담에 있어서는 의사보다 공감을 더 잘해주는 것으로 나타났다.[11]

b. Medical Information Retrieval

앞서 기술한 바와 같이, LLM 의 본래 기능은 대화가 아니며 주어진 텍스트의 빈칸 채우기 (e.g., BERT, T5) 혹은 다음 단어의 예측이다 (GPT, LLaMa). 그러므로, 자유 형식의 의무기록과 함께 필요한 서식을 함께 프롬프트에 넣으면, 원하는 형태로 기록을 가공하거나 정리, 요약 및 특정 정보를 추출하는데 이용할 수 있다.[12,13] 특히, 정확한 포맷을 철저히 지키도록 유도할 수 있기 때문에 생성된 텍스트로 다른 프로그램을 직접적으로 조작할 수 있게 된다.[14] 일례로 수학, 산수 문제는 LLM 의 대표적인 약점으로 알려져 있는데 이로 인해 Liver Imaging Reporting & Data System (LI-RADS)과 같이 영상 소견을 세고, 크기를 비교해야 하는 작업의 정확도가 떨어진다. 이러한 문제를 외부 프로그램(Python code 등)을 결합하여 원본 의무기록에서의 feature extraction 만 LLM 에게 맡기고, categorization 자체는 기계적으로 수행하게 된다면 상당히 높은 수준의 정확도를 확보할 수 있다.[13] 유사한 연구가 폐암 환자들의 흉부 CT 판독문에 대해서도 수행된 바 있고, 앞으로 더 고도화된 LLM 을 통해 영상의학 전 분야에서 판독문에 대한 신뢰도 높은 데이터 마이닝이 손쉽게 가능해질 수 있을 전망이다.[15]

References

1. Bhayana R, Bleakney RR, Krishna S. GPT-4 in Radiology: Improvements in Advanced Reasoning. *Radiology* 2023:230987
2. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* 2023
3. Dolfing HJ, Hetherington IL. Incremental language models for speech recognition using finite-state transducers. In:*IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*: IEEE, 2001; 194-197
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in neural information processing systems* 2017;30
5. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog* 2019;1:9
6. Chen M, Tworek J, Jun H, Yuan Q, Pinto HPdO, Kaplan J, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* 2021
7. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns* 2023
8. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* 2023
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180
10. Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, Nadler D, et al. Do we still need clinical language models? In:*Conference on Health, Inference, and Learning*: PMLR, 2023; 578-597
11. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 2023;183:589-596
12. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023:230725
13. Gu K, Lee JH, Shin J, Hwang JA, Min JH, Jeong WK, et al. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. *Liver International* 2024:In press
14. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, et al. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 2024;36
15. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 2023;308:e231362