LLMs are key components to be able to extract unstructured information in clinical settings. Some articles discussed different ways that are effective in retrieving this information.

The article, “Large Language Models are Few-Shot Clinical Information Extractors”, focused on the more general purpose of LLMs and if they can perform information extraction with few-shot learning. Supervised deep learning performed well when annotated data was available and pretrained biomedical language models improved generalization compared to general language models. Using annotated corporas like i2b2 as a standardized benchmark and having carefully designed schemas also improve performance for named entity recognition (NER). There are limitations as labeled datasets are scarce and fine tuning costs are expensive.

- https://arxiv.org/pdf/2205.12689

In the article, “Large Language Models for More Efficient Reporting of Hospital Quality Measures”, the focus was on automating hospital quality reporting more specifically severe sepsis and septic shock (SEP-1). A hybrid approach of using structured data and NLP receives good accuracy. Also, in standardized EHRs with consistent terminology, rule based pipelines were effective. The article uses human abstraction as the benchmark and automation to assist in pre screening. This method did not work well for unstructured text and the models could not distinguish what content was relevant. Also, automation did not generalize and the manual review used for validation was costly.

- https://ai.nejm.org/doi/full/10.1056/AIcs2400420

The article, “GPT-NER: Named Entity Recognition via Large Language Models”, focused more on improving NER through LLMs like gpt-3 in few-shot settings. The article found that using LLMs as text generators allows for adaptation to NER tasks without having to retrain. They used text generation with entity markers and double checked predictions to reduce false positives. This achieved better F1 scores compared to the supervised baselines, and the LLMs generalized well to different datasets and nested NER tasks. Some difficulties that occurred was that the LLMs were still prone to hallucinations for domain specific tasks, it is largely dependent on prompt design, and has high cost compared to other NER models.

- https://arxiv.org/pdf/2304.10428

In the article, “Zero-Shot Clinical Trial Patient Matching with LLMs”, they focused on being able to match patients to clinical trials by using LLMs as an information extraction (IE) task. Supervised IE models, rule based systems with standardized text, and domain specific datasets allow for good performance showing that it is mostly effective for structured matching when data is clean and available. Thus, this approach has limited generalization and a bottleneck as labeling the data is costly and time consuming.

- https://arxiv.org/pdf/2402.05125

Another article called “Health system-scale language models are all-purpose prediction engines”, uses unstructured clinical notes in an LLM to create a flexible predictive model as existing models rely on structured EHR data making it difficult to deploy in real world situations. The article found that pretraining then fine tuning the model on a large amount of unstructured clinical data worked well. Also, that using unstructured text outperforms structured data baselines for certain tasks. This led them to conclude that real world deployment is possible as domain specific pretraining improves efficiency and allows generalization. Some limitations that arise are that the LLMs are not zero shot and still need fine tuning, and although generalization occurs it is dependent on the hospital/department.

- https://www.nature.com/articles/s41586-023-06160-y