Overall, LLMs are good at extracting unstructured information from EHR records. For simple extraction of information, we can use anti-hallucination prompt strategies and can achieve good performance. Few shot inference, where we give a few examples to the LLMs, also consistently is shown to help performance. For more complicated tasks like symptom labelling, then a technique like RAG is often more powerful.

I've looked at these articles more in-depth:
https://pmc.ncbi.nlm.nih.gov/articles/PMC11751965/: This 2025 study found that LLMs were very effective in general at extracting EHR information from 50 synthetic test cases. Claude models were the best. This shows potential for LLMs in the exact domain that we need

https://pmc.ncbi.nlm.nih.gov/articles/PMC12099322/: This 2024 study used LLMs to extract numerical, unstructured data into a structured format instead of using regex-based approaches. Llama is open-source, so it helps with data security issues. They used several heuristics for post-processing, for example putting a range of values that are accurate to extract information. Overall, the outputs were about as effective as regex approaches and requires significantly less time from a user to carefully craft a regex – thus llms show promise.

https://www.nature.com/articles/s41598-025-00724-w: This 2025 paper used RAG to help LLMs give good answers for diagnosis. In order to make the RAG more effective, they used a more clever dual-store retrieval and went beyond pure semantic search. Overall, RAG is useful if we want more complex information retrieval from common data stores like in this paper.

I've also seen some papers on fine-tuning etc. but did not look too closely into them.