Informal Literature Review for LLM as Info Extractor

Kaijie Zhang

In most papers of 'LLM based diagnostic information extraction in clinical text', improvement methods mainly revolve around reducing hallucination. This reduces a complex problem to classic LLM issues, but EHR settings are more stringent. Incorrect extraction (hallucination or incorrect format) can lead to serious clinical consequences.

Mainstream methods and papers

We can see many commonly used LLM techniques applied here, and most of them require extra EHR specific adaptation.

Prompt based learning and chain of thought (CoT): Through carefully designed prompts and chain-of-thought reasoning, diagnostic information is extracted directly from the text. For example: [https://www.nature.com/articles/s41746-024-01377-1].

Fine tuning and multi task learning: It is very less likely for us to fine tune any LLMs due to limited computing resources. But you can read [https://arxiv.org/abs/2411.15700] for this.

Retrieval augmented generation (RAG): We should be familiar with it. Read the first paper I gave for reference.

Knowledge and schema aware methods: Some works incorporate domain knowledge bases and staged processes to enhance accuracy. This is highly HER-oriented. Read [https://arxiv.org/html/2406.18027v1].

Data augmentation and model architecture: In addition to traditional techniques, to address the challenges of limited annotations and long texts, the industry has proposed LLM-assisted data augmentation and new architectural designs. We can think about that. Read [https://pubmed.ncbi.nlm.nih.gov/40776002/].

SOTA models and metrics

Model performance often varies with data. For complex clinical variable extraction, the CLEAR experiments on real world notes reported an average F1 of 0.90 nature.com. On the Stanford MOUD dataset, GPT 4 achieved F1 of 1.00 on ten variables with an average F1 of 0.90, which can serve as a reference. This performance is from the first paper I shared above [https://www.nature.com/articles/s41746-024-01377-1].

We may pay more attention to these metrics:

Accuracy and recall improvement

Long document understanding and context integration

Few shot ability (I am not sure how much data we can access for now, but rare diseases require robustness to few shot settings)

We will use ablation studies to observe performance changes, so this part is not a major concern for now.

Challenges

Most papers address hallucination, but their targets differ. The following challenges still have considerable room for improvement.

Hallucination and accuracy. LLMs tend to produce hallucination, that is diagnostic information that does not appear in the source report. Because LLM training data are static and largely general domain, their understanding of domain specific terminology can be limited, and they often return incorrect diagnoses or extraneous information. Although retrieval augmented generation can mitigate this to some extent, we still need more reliable mechanisms to ensure that extracted fields are factually grounded and trustworthy.

Consistency and structure. When an LLM is asked to output structured fields directly, the results often lack consistency and canonical formatting. For multi-level nested information, for example multiple lesions and their attributes at multiple time points, LLMs often ignore secondary information or produce messy formats. Two stage extraction with schema validation improves this locally, but end to end structural consistency remains difficult.

Context drift and domain updates. Clinical knowledge and practice evolve, and the static knowledge in an LLM can lag behind. Writing styles and terminology vary across institutions and departments, which can cause performance degradation. How to enable continual adaptation or better use of up-to-date medical knowledge remains an open problem.

Multimodal fusion. Whether we consider this depends on the data we can access at UCSD Health. Most current work focuses on text only, while real diagnostic information often spans modalities such as imaging, waveforms, and pathology. Some studies begin to combine images and text, for example GPT 4V style vision and language on radiology images and reports, but how to unify multimodal features and ensure correct diagnostic reasoning is still an open challenge.

New diseases and rare conditions. For newly emerging diseases or rare conditions, due to the lack of clear annotation or precedent, an LLM may fail to recognize them. Chain of thought and self-consistency can increase hit rates on rare diagnoses, for example, check [https://www.nature.com/articles/s44387-025-00011-z]. But truly label free zero shot extraction for novel conditions is still immature, and future work needs medical knowledge bases and dynamic learning to improve this.

Deployment and API integration

If we have bandwidth beyond research, I suggest we land the system in practice. Many studies have done similar things. See example [https://www.nature.com/articles/s41698-025-01103-4].