Exposé - Study Project
Evaluating Large Language Models based on Clinical Practice Guidelines for Precision Oncology

Ayda Yilmazer
Superviser: Xing David Wang

Humboldt-Universität zu Berlin

July 5, 2024

1 Introduction

In the past decade, Large Language Models (LLMs) have managed to attract considerable interest. They have the potential to be utilized across various applications, from education to entertainment, finance, research and engineering [1].

LLMs can also be used to gather information in medicine for understanding symptoms, learning about medical conditions, or seeking general health advice. A recent study explored how LLMs are increasingly being used by the public to obtain medical information [2].

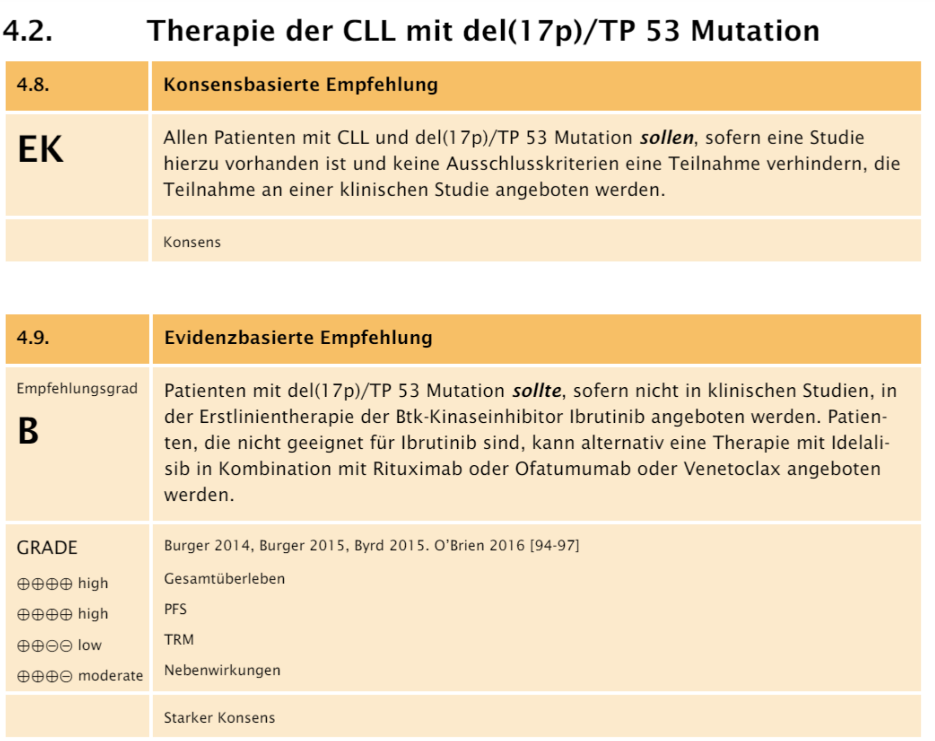
Another potential application field is precision oncology. In precision oncology, treatment is customized based on the specific molecular profile of each patient's tumor. The assessment of possible treatment options occurs in molecular tumor boards (MTBs). An MTB is an interdisciplinary group of medical experts, whose goal is to analyze clinical and molecular profiles of cancer patients with rare or advanced tumors to identify additional therapy options, in case the treatment options based on already established guidelines are no longer viable. The cases that are discussed in MTBs are complex, requiring intricate analysis of molecular profiles of patients by multiple medical experts [3]. This analysis is an elaborate process and takes place in form of a manual search through large text-based resources and knowledge bases, and is considered a bottleneck of precision oncology [4].

LLMs are known to excel at various text-based tasks, including leveraging the extensive knowledge stored in large amounts of text data and using that knowledge for decision support and problem solving in various scenarios [5]. Multiple studies have been conducted to explore the ways to utilize LLMs to enhance clinical decision support (CDS), including precision oncology scenarios [6]. These studies conclude that the performance of LLMs in clinical decision making falls short of the desired quality for reasons like the following:

- Human experts did not completely agree with the recommendations made by the LLMs [7].
- The LLMs gave ambiguous outputs [8].
- Some treatment recommendations were hallucinated and some were not in line with the guidelines [8].

In this study project, we aim to assess the capabilities of two LLMs', LLama 3 [9] and ChatGPT 4 [10], in providing useful treatment recommendations for cancer patients and determine how consistent the recommendations are with the standard treatment recommendations. Therefore we are looking at Clinical Practise Guidelines (CPGs) for common mutational profiles in cancer to assess the LLMs' performances for such cases.

CPGs are documents that contain recommendations on diagnosis, treatment and follow-up care of a medical condition [11]. They are aimed at healthcare professionals and provide a summary of the current knowledge of the condition. Usually, they are available in PDF or HTML format. One exemplary cancer guideline is the guideline for chronic lymphocytic leukemia by the German Oncological Society Leitlinienprogramm Onkologie [12]. Figure 1 contains a snippet of the above-mentioned guideline and serves as an example of how a common mutation is handled in a cancer guideline.



4.2. Therapie der CLL mit del(17p)/TP 53 Mutation

4.8.	Konsensbasierte Empfehlung
EK	Allen Patienten mit CLL und del(17p)/TP 53 Mutation ***sollen***, sofern eine Studie hierzu vorhanden ist und keine Ausschlusskriterien eine Teilnahme verhindern, die Teilnahme an einer klinischen Studie angeboten werden.
	Konsens

4.9.	Evidenzbasierte Empfehlung
Empfehlungsgrad B	Patienten mit del(17p)/TP 53 Mutation ***sollte***, sofern nicht in klinischen Studien, in der Erstlinientherapie der Btk-Kinaseinhibitor Ibrutinib angeboten werden. Patienten, die nicht geeignet für Ibrutinib sind, kann alternativ eine Therapie mit Idelalisib in Kombination mit Rituximab oder Ofatumumab oder Venetoclax angeboten werden.
GRADE	Burger 2014, Burger 2015, Byrd 2015. O'Brien 2016 [94-97]
⊕⊕⊕⊕ high	Gesamtüberleben
⊕⊕⊕⊕ high	PFS
⊕⊕⊖⊖ low	TRM
⊕⊕⊕⊖ moderate	Nebenwirkungen
	Starker Konsens

Figure 1: A snippet from the guideline for chronic lymphocytic leukemia by the German Oncological Society Leitlinienprogramm Onkologie [12], showing an evidence-based treatment recommendation for the del(17p)/TP 53, a common mutation [13] of chronic lymphocytic leukemia

2 Related Work

The potential uses of LLMs for CDS, including precision oncology, have already been explored in numerous studies. It is widely assumed that the integration of LLMs has significant potential to enhance clinical decision-making during the diagnosis and prognosis stages, leading to an improved treatment for oncology patients.

Zhou et al. [14] conducted a study to compare the performance of ChatGPT-3.5, ChatGPT-4, ChatGPT-4 Turbo, Doctor GPT, LLaMa-2-70B, Mixtral-8x7B, Bard (Gemini Pro), and Claude 2.1 according to 9 oncology physicians (three residents, three fellows, and three attendings) in answering questions related to colorectal cancer. They used the National Comprehensive Cancer Network (NCCN) guidelines for colon and rectal cancer to generate 150 close-ended questions. The authors evaluated the answers of the LLMs and the physicians on their consistency with the NCCN guidelines. They reported that for all LLMs, a higher proportion of the answers were concordant with the NCCN guidelines. The LLMs that had the greatest accuracy were Claude 2.1 with approx. 83%, Doctor GPT with approx. 80% and ChatGPT-4 Turbo with approx. 78%. Claude 2.1 outperformed both fellows and attendings, Doctor GPT outperformed the fellows and Mixtral-8x7B, ChatGPT-4, and ChatGPT-4 Turbo outperformed the residents. They stated that their findings show that these chatbots are capable of providing correct medical information on colorectal cancer, and that LLM-powered chatbots could play a role in tumor boards. The authors also observed that there was an issue with hallucinated answers with all of the LLMs, but especially with Claude 2.1. Claude 2.1 has answered over 95% of the questions confidently, whereas other chatbots answered less than 50%. Furthermore, all LLMs except LLaMa-2-70B outperformed Claude 2.1 by achieving over 90% accuracy in their confident replies. They stated that hallucinations are the main obstacle to the development of AI and that LLM chatbots should not be relied on and used without human expert reviews.

Chen et al. [8] examined how well ChatGPT's suggested courses of treatment for prostate, lung, and breast cancer correspond with the NCCN guidelines. They created four distinct prompt templates for querying the chatbot about treatment suggestions given 26 synthetic diagnosis descriptions. The prompts were zero-shot, i.e., they did not contain any guideline-related information. They concluded that the chatbot could not be considered a trustworthy source of treatment information due to the inaccuracy of the recommendations it offered. One-third of the recommendations made by the LLM were partially inconsistent with the National Comprehensive Cancer Network guidelines. They also observed that the ambiguous output of the chatbot caused conflicts among the annotators.

Sorin et al. [7] investigated ChatGPT as a support tool for a breast tumor board. They compared the tumor board's and ChatGPT's suggestions for ten cancer patients, based on clinical information of the patients. They concluded that in 70% of the cases, the chatbot's suggestions coincided with those of the tumor board. The chatbot's performance in summarization and explanation were however rated higher by the reviewers. They also observed certain problems, such as the chatbot disregarding information about a patient's test results and the lack of a recommendation for a consultation with a radiologist.

Similarly, Benary et al. [6] assessed 4 LLMs as support tools for precision oncology. They created 10 fictional patients with advanced cancer and submitted the cases to 4 recent LLMs and 1 expert physician to suggest treatment options. They then presented the proposed treatment options to a MTB and let the members assess the probability of a treatment option being suggested by a LLM and whether the treatment option was beneficial. They reported that, while the LLMs' treatment suggestions did not meet the standards of human experts, one LLM generated two useful treatment options that the human experts could not identify.

Oniani et al. [15] have researched incorporating Centers for Disease Control and Prevention and Infectious Diseases Society of America COVID-19 Treatment Guidelines into LLMs to enhance CDS. They created three distinctive methods (Binary Decision Tree, Program-Aided Graph Construction and Chain-of-Thought-Few-Shot Prompting) and used a collection of synthetic patient profiles to evaluate the responses in two stages. In the first stage they evaluate the methods using the F-Score metric and select those with a score greater than 0.5. In the second stage, two physicians rated the responses of the LLMs in three different evaluation categories set by the authors: presence of incorrect medical content, omission of content, and presence of possible harmful content. The first category's goal is to identify if an answer contains information that conflicts medical guidelines. The second category aims at evaluating the completeness of the answers. Lastly, with the third category, the authors' goal is to determine whether an answer contains information that might cause harm to the users. In addition to the three methods they developed, the authors queried the LLMs with zero-shot prompts, which served as a baseline for the evaluation. As a baseline, they used zero-shot prompting. They found that the LLMs enhanced with CPGs outperformed the LLM with zero-shot prompting. They claim that to their knowledge, their study is the first one to develop and evaluate methods to augment LLMs with CPGs. Thus, this study demonstrates that it is possible to improve LLM decision-making performance using in-context learning methods.

3 Goal of the Study Project

In this work, we plan to evaluate the LLMs Llama 3 and ChatGPT 4 in providing useful treatment suggestions for patients with various types of cancer and common molecular profiles. We will evaluate the treatment suggestions based on how well they align with established CPGs.

To achieve this, we will first create a data set containing profiles of various fictional cancer patients using CPGs. Next, we will develop a prompt template to ask the LLMs to suggest treatment strategies for the patients. We will then evaluate the answers given to the prompts by the LLMs.

We stress that the goal of the study project is not to create a a sophisticated prompting guide but first to establish a foundational data set based on clinical guidelines for future research to build upon.

4 Approach

The study project will include the following steps:

4.1 Data Extraction

The first task of this project will be to create the fictional patient profiles. For this, we first need to decide which information to include before starting the actual data creation process. This information may include comorbidities, disease-specific information like mutational profile, previous treatments, etc. Selecting the appropriate attributes is crucial, since the LLMs will use this information to give their recommendations. This information will be manually extracted from several various CPGs for oncology. We will consider 10-20 types of cancer for the patient profiles and all of the profiles will consist of the same attributes. Additionally, we will add to the profiles redundant attributes that will operate as noise and that the LLMs will not necessarily need to consider when making a treatment suggestion, such as diet, occupation and race. With the latter, we can also investigate whether the LLMs have racial prejudice.

Once the attributes of a patient profile have been determined, we will create the patient data using the CPGs. We will manually create fictional patient profiles using the information provided in the CPGs for various cancer types. All created data will be unified and structured so that it fits into the prompt template (see Section 4.3), resulting in a natural-sounding and unambiguous prompt.

Figure 2 shows a detailed overview of the Data Extraction step. In Step 1, we will go through various guidelines and choose the ones that are more refined and consist of established standards. In Step 2, we will use the available information in the guidelines to determine the attributes the patient profiles will consist of. In this step, we will consider all of the guidelines we have chosen in Step 1 to ensure that the attributes are the same for every patient. Then, in Step 3 we will go through the guidelines comprehensively to collect the information needed for the patient profiles. In contrary to Step 2, this time, we will look at the guidelines separately for every patient. Lastly, in Step 4 we will adjust the collected information to fit into the prompt template seamlessly.

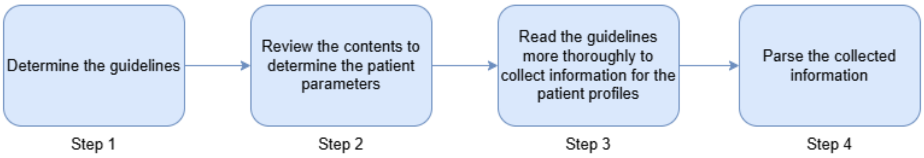

Determine the guidelines
Step 1
Review the contents to determine the patient parameters
Step 2
Read the guidelines more thoroughly to collect information for the patient profiles
Step 3
Parse the collected information
Step 4

Figure 2: Flowchart for the workflow in the Data Extraction step (Section 4.1)

4.2 Clinical Practise Guidelines Dataset

In this project, we will focus on the German CPGs. We will scrape the guidelines and parse the data manually to determine the attributes and collect the information for the patient profiles.

We will use the CPGs from Leitlinienprogramm Onkologie [16] and Onkopedia [17] in our project. Leitlinienprogramm Onkologie is a platform by Deutsche Krebsgesellschaft e.V. and provides cancer guidelines in PDF format, whereas Onkopedia Guidelines, a project by Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. consists of guidelines in both PDF and HTML format.

4.3 Creating the Prompt Template

To interact with the LLMs, we will define fixed prompt templates. The LLMs will be asked for suggestions using these prompts. The prompts will include placeholders that will be filled with the patient information. Our goal while developing the prompt templates is to ensure that the LLMs do not overlook any information regarding the patients. Table 1 shows two example prompts that we intend to use in this project. Prompt 1 is a natural language question, whereas Prompt 2 provides the information in a more structured way. In this project, we will only be focusing on zero-shot prompts. We are not going to share any information about the guidelines in our prompt, nor will we fine-tune the models.

Prompt 1	Given a <*gender*> patient with age <*age*> and <*previous diseases (or no disease history)*> diagnosed with <*diagnosis*> stage <*stage*>, mutations <*mutations, if any*>, what are the possible treatment options?
Prompt 2	Recommend targeted treatment options for the following patient: Gender: <*gender*>, Age: <*age*>, Disease History: <*previous diseases (or no disease history)*>, Diagnosis: <*diagnosis*> Stage <*stage*> with mutations <*mutations, if any*>>

Table 1: Two example prompt templates.

4.4 Evaluation

After querying the LLMs using the created patient profiles and prompts, we will evaluate the answers. Because we are using CPGs to create patient data, we will have access to the standard treatment recommendations in the guidelines. We will save this information in a data set along with the related patient data and use it to assess the LLMs' responses. We will evaluate the performances compared with the standard treatments in the guidelines using the F1-Score metric:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Precision denotes the fraction of relevant treatment suggestions among all suggestions made by that LLM and is defined as:

$$precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall denotes the fraction of relevant treatment options found by the LLMs and is defined as:

$$recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

During the comparison, we will watch out for possible discrepancies in drug nomenclature between the LLMs and the guidelines, where the LLMs may refer to a drug by using a name different from the one specified in the guidelines, to ensure an accurate evaluation.

References

[1] Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage, July 2023.

[2] Yunpeng Xiao, Kyrie Zhixuan Zhou, Yueqing Liang, and Kai Shu. Understanding the concerns and choices of public when using large language models for healthcare, January 2024. arXiv:2401.09090 [cs].

[3] Mario Lamping, Manuela Benary, Serge Leyvraz, Clemens Messerschmidt, Eric Blanc, Thomas Kessler, Moritz Schütte, Dido Lenze, Korinna Jöhrens, Susen Burock, Konrad Klinghammer, Sebastian Ochsenreither, Christine Sers, Reinhold Schäfer, Ingeborg Tinhofer, Dieter Beule, Frederick Klauschen, Marie-Laure Yaspo, Ulrich Keilholz, and Damian T. Rieke. Support of a molecular tumour board by an evidence-based decision management system for precision oncology. *European Journal of Cancer*, 127:41–51, March 2020.

[4] Benjamin M. Good, Benjamin J. Ainscough, Josh F. McMichael, Andrew I. Su, and Obi L. Griffith. Organizing knowledge to enable personalization of medicine in cancer. *Genome Biology*, 15(8):438, August 2014.

[5] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):160:1–160:32, April 2024.

[6] Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Network Open*, 6(11):e2343689, November 2023.

[7] Vera Sorin, Eyal Klang, Miri Sklair-Levy, Israel Cohen, Douglas B. Zippel, Nora Balint Lahat, Eli Konen, and Yiftach Barash. Large language model (ChatGPT) as a support tool for breast tumor board. *npj Breast Cancer*, 9(1):1–4, May 2023. Publisher: Nature Publishing Group.

[8] Shan Chen, Benjamin H. Kann, Michael B. Foote, Hugo J. W. L. Aerts, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncology*, 9(10):1459–1462, October 2023.

[9] Meta. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`. Accessed: 2024-05-16.

[10] OpenAI. Introducing chatgpt. `https://openai.com/index/chatgpt/`. Accessed: 2024-05-16.

[11] Institute for Quality and Efficiency in Health Care (IQWiG). In brief: What are clinical practice guidelines? `https://www.ncbi.nlm.nih.gov/books/NBK390308/`. Accessed: 2024-05-14.

[12] Leitlinienprogramm Onkologie. S3-Leitlinie chronisch lympatische Leukämie (CLL). `https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/CLL/LL_CLL_Langversion_1.0.pdf`. Accessed: 2024-05-14.

[13] Maria de Lourdes L.F. Chauffaille, Ilana Zalcberg, Wolney Gois Barreto, and Israel Bendit. Detection of somatic TP53 mutations and 17p deletions in patients with chronic lymphocytic leukemia: a review of the current methods. *Hematology, Transfusion and Cell Therapy*, 42(3):261–268, 2020.

[14] Shan Zhou, Xiao Luo, Chan Chen, Hong Jiang, Chun Yang, Guanghui Ran, Juan Yu, and Chengliang Yin. The performance of large language model powered chatbots compared to oncology physicians on colorectal cancer queries. *International Journal of Surgery*, page 10.1097/JS9.0000000000001850.

[15] David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. Enhancing Large Language Models for Clinical Decision Support by Incorporating Clinical Practice Guidelines, January 2024. arXiv:2401.11120 [cs].

[16] Deutsche Krebsgesellschaft e.V. Leitlinienprogramm onkologie. `https://www.leitlinienprogramm-onkologie.de/leitlinien/uebersicht`. Accessed: 2024-06-13.

[17] Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. Onkopedia guidelines. `https://www.onkopedia.com/de/onkopedia/guidelines`. Accessed: 2024-06-13.