Article published online: 2024-02-26

Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning

Verbesserung des Einsatzes von Großen Sprachmodellen in der Radiologie durch „Prompt Engineering“: von präzisen Prompts zu Zero-Shot Learning

Authors
Maximilian Frederik Russe, Marco Reisert, Fabian Bamberg, Alexander Rau

Affiliations
Department of Radiology, University Hospital Freiburg, Freiburg, Germany

Keywords
radiology, artificial Intelligence, natural Language Processing

received 24.10.2023
accepted 30.1.2024
published online 2024

Bibliography
Fortschr Röntgenstr
DOI 10.1055/a-2264-5631
ISSN 1438-9029
© 2024. Thieme. All rights reserved.
Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence
Dr. Alexander Rau
Department of Diagnostic and Interventional Radiology, Universitätsklinikum Freiburg Klinik für Radiologie, Breisacher Straße 64, 79106 Freiburg, Germany
Tel.: +49 761 270 38190
alexander.rau@uniklinik-freiburg.de

ABSTRACT

Purpose Large language models (LLMs) such as ChatGPT have shown significant potential in radiology. Their effectiveness often depends on prompt engineering, which optimizes the interaction with the chatbot for accurate results. Here, we highlight the critical role of prompt engineering in tailoring the LLMs’ responses to specific medical tasks.

Materials and Methods Using a clinical case, we elucidate different prompting strategies to adapt the LLM ChatGPT using GPT4 to new tasks without additional training of the base model. These approaches range from precision prompts to advanced in-context methods such as few-shot and zero-shot learning. Additionally, the significance of embeddings, which serve as a data representation technique, is discussed.

Results Prompt engineering substantially improved and focused the chatbot’s output. Moreover, embedding of specialized knowledge allows for more transparent insight into the model’s decision-making and thus enhances trust.

Conclusion Despite certain challenges, prompt engineering plays a pivotal role in harnessing the potential of LLMs for specialized tasks in the medical domain, particularly radiology. As LLMs continue to evolve, techniques like few-shot learning, zero-shot learning, and embedding-based retrieval mechanisms will become indispensable in delivering tailored outputs.

Key Points

- Large language models might impact radiological practice and decision-masking.
- However, implementation and performance are dependent on the assigned task.
- Optimization of prompting strategies can substantially improve model performance.
- Strategies for prompt engineering range from precision prompts to zero-shot learning.

ZUSAMMENFASSUNG

Ziel Große Sprachmodelle (engl. LLMs) wie ChatGPT haben ein erhebliches Potenzial in der Radiologie gezeigt. Ihre Effektivität hängt oft vom sog. Prompt-Engineering ab, das die Interaktion mit der künstlichen Intelligenz für genaue Ergebnisse optimiert. Hier wird die kritische Rolle des Prompt-Engineerings bei der Anpassung der Antworten der LLMs an spezifische medizinische Aufgaben hervorgehoben.

Material und Methoden Anhand eines klinischen Falles erläutern wir verschiedene Prompting-Strategien zur Anpassung des LLM ChatGPT mit GPT4 an neue Aufgaben ohne zusätzliches Training des Basismodells. Diese Ansätze reichen von präzisierten Prompts bis hin zu fortgeschrittenen In-Kontext-Methoden wie „few-shot“- und „zero-shot“-Lernen. Zusätzlich wird die Bedeutung des „Embeddings“ als Datenrepräsentationstechnik diskutiert.

Ergebnisse Das Prompt-Engineering verbesserte und fokussierte den Output des Chatbots erheblich. Darüber hinaus ermöglicht die Einbettung von Fachwissen einen transparenteren Einblick in die Entscheidungsfindung des Modells und stärkt so das Vertrauen.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Schlussfolgerung Trotz gewisser Herausforderungen spielt das Prompt-Engineering eine zentrale Rolle bei der Nutzung des Potenzials von LLMs für spezialisierte Aufgaben im medizinischen Bereich, insbesondere in der Radiologie. Im Zuge der Weiterentwicklung von LLMs werden Techniken wie „few-shot learning“, „zero-shot learning“ und „Embedding“ für die Bereitstellung maßgeschneiderter Ergebnisse unverzichtbar werden.

Kernaussagen

- Große Sprachmodelle könnten die radiologische Routine und Entscheidungsfindung beeinflussen.
- Die Implementierung und Leistung hängen jedoch von der zugewiesenen Aufgabe ab.
- Die Optimierung von Prompting-Strategien kann die Modellleistung erheblich verbessern.
- Strategien für das Prompt-Engineering reichen von Präzision-Prompts bis zum Zero-Shot-Lernen.

Zitierweise

- Russe MF, Reisert M, Bamberg F et al. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning . Fortschr Röntgenstr 2024; DOI 10.1055/a-2264-5631

Main Body

Human interaction is primarily governed by language, a system traditionally reserved for human use due to its complexity. Recent advances in algorithms based on the transformer architecture, however, leveraged artificial intelligence (AI) to understand and generate human-like text, especially those using large language models (LLMs). These models have the potential to revolutionize communication by processing human language. A recent breakthrough in this regard was the release of the chatbot ChatGPT into broad use by OpenAI in November 2022 [1]. While its recreational applications are notable, its potential extends well beyond entertainment. It has shown considerable promise in data-intensive fields, particularly in medicine and especially within radiology.

Applications include passing the United States Medical Licensing Examination [2], simplifying radiology reports [3–6], extracting information from oncology notes [7], and classifying fractures based on free-text reports [8].

A particular challenge in the interaction with LLMs like general conversational chatbots is to focus the model on the particular task. This guidance is done by means of the input – the so-called prompt. The selection of the prompt to the model has great influence on whether the problem is processed in the targeted manner. Prompt engineering optimizes this interaction [9, 10].

For example, instead of a generic question such as “Tell me about atypical pneumonia”, a more nuanced prompt might be “Provide a detailed explanation of atypical pneumonia as opposed to typical pneumonia findings on CT scans”. In addition, role-based prompting can be used to encourage the model to generate responses from a particular perspective, like that of an expert or novice, e. g., “You are ChatGPT a subspecialized thoracic radiologist”. Furthermore, audience-aware prompts can guide the model to tailor the response based on the intended recipients, whether they are specialists in a field or laypeople. For example, the latter would require a prompt like “Give your answer in easy language” to reduce the complexity and medical jargon. These adaptations add layers of customization that ensure more targeted and useful output [3].

LLMs can be designed to recognize and parse free-text radiology reports, extracting key findings, measurements, and anatomical regions. For example, if a report mentions “a 3.5 cm mass in the upper lobe of the right lung,” the LLM can categorize this as a “finding” with specific attributes (location, size, etc.) Once the information is extracted, LLMs can automatically populate structured report templates, and even be used to accelerate the development of other non-language-based deep learning approaches [11].

As foundational as prompt engineering is, it sets the stage for more adaptive methodologies like few-shot learning, where the AI’s adaptability is challenged beyond structured guidance.

Reminiscent of traditional supervised learning, few-shot learning allows for an adaptive approach in the context of the question. While prompt engineering relies primarily on guiding the model with meticulously crafted queries, few-shot learning allows the model to adapt to new tasks, even with a very limited set of examples [12, 13].

Imagine a scenario in which a medical AI system receives a series of structured examples such as “Calcified nodule = Benign”, “Pneumonia = Benign”, “Spiculated mass = Malignant”. After these few examples, using few-shot learning principles, the system might deduce the response of the query “Infiltrating tumor = ?” to be “Malignant”. This decision is made not only based on extensive knowledge about radiological findings (incorporated in the general training data) but by recognizing patterns in the provided examples and extrapolating from them.

Such an approach is paramount to the medical world, especially when dealing with rare diseases or novel treatments in which large amounts of data are lacking. Instead of requiring an immense amount of data or operating in the absence of direct experience, few-shot learning leverages these handful of examples to generate insight.

Zero-shot learning further extends this approach and is a pivotal advancement of artificial intelligence, especially in the context of LLM. In contrast to traditional machine learning paradigms, zero-shot learning empowers models to execute tasks they haven’t been directly trained for. Rather than relying on specific labeled data for every task, these models leverage their vast general knowledge and specific context provided in the prompt for output generation [13, 14].

Such an approach could have been employed in the early days of the COVID-19 outbreak, when knowledge about the virus was scarce. A medical professional could present a detailed description of a patient’s symptoms and medical history to an LLM. Together with the patient-related information, specialized knowl-

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

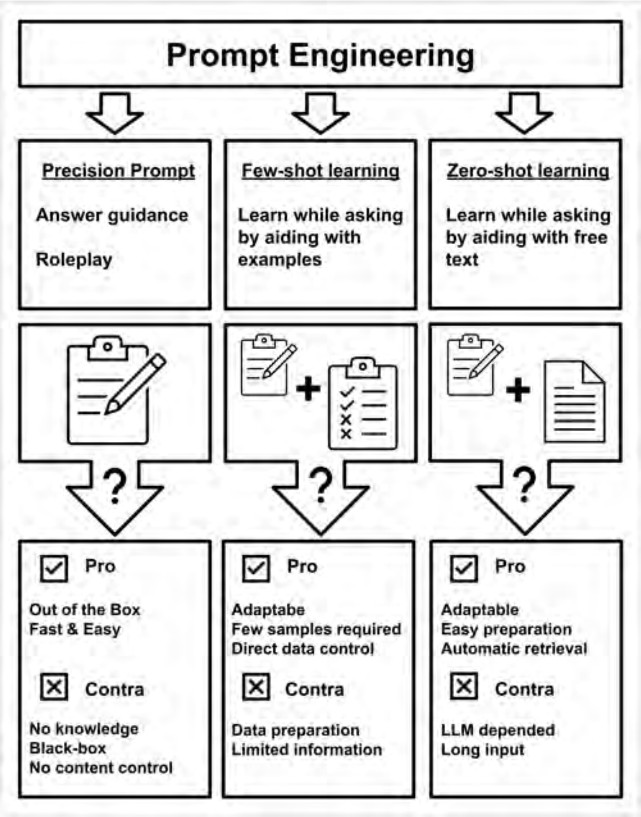
edge would be introduced with free-text information from a recent research paper detailing the symptoms of the emerging virus. In response to the query “Does this patient have symptoms consistent with COVID-19?”, the LLM could cross-reference the provided information and recognize patterns consistent with the described symptoms of the virus – even without explicit prior training or data on COVID-19-related information. Using this zero-shot learning, the LLM might respond as follows: “The patient’s symptoms align closely with those described for COVID-19 in the provided research”. The ability to make informed decisions in the absence of direct task-related training sets zero-shot learning apart from traditional learning techniques [13].

Embeddings are fundamental to enable artificial intelligence to automatically retrieve content for zero-shot learning instead of manually selecting the data. These embeddings are a form of data representation, where complex data such as text, is converted into vectors in a multi-dimensional space. These numerical vectors capture the context, semantics, and relationships inherent in the original data [15]. This leads to an alternative approach of retrieving relevant content: Instead of searching through data using traditional keyword-based methods, it employs the vectorized embeddings to reveal related content by mathematically deducing the nearest neighbor in the semantic meaning. In medical settings, the implications are profound. By converting a patient’s symptoms and medical history into embeddings, the system could browse a vector library to retrieve related articles, research papers, or even similar patient cases [7, 8, 16]. Embeddings thus can ensure that the LLM receives the most relevant and contextually appropriate information, allowing for superior output creation. Moreover, the accessed information in the vector library can be curated and easily updated to the local needs. With respect to the growing volume of medical data, such vector-based retrieval mechanisms will become indispensable tools for both medical professionals and researchers.

An overview of these strategies is given in ▶ **Fig. 1** and the respective potential to optimize output is illustrated with a clinical example case requiring imaging in ▶ **Table 1**. Please note that only the precision prompt recommended an unnecessary procedure (non-contrast CT), which highlights the potential of more advanced prompting strategies.

In general, prompting an LLM in a radiological setting might comprise the following:

1. Task and Role: Concisely name the task and define the intended role of the LLM.
2. Contextual Setup (if applicable): Provide relevant context or the background information necessary for the task enabling few/zero-shot learning.
3. Task-Specific Instructions: Give concise, direct instructions related to the specific task at hand.
4. Output Format Specification: Define how you want the output to be structured. This can include the level of detail, the format of the response, or any specific points or structure that need to be addressed.
5. Cautionary Note: Include any cautionary notes to prevent overcomplication or misunderstanding of the task.



▶ **Fig. 1** Schematic overview of basic and advanced prompting strategies.

6. User Data Input Specification: Clearly specify what data or information is being input, whether it’s textual descriptions, image summaries, patient history, etc.

For example, a precision prompt to simplify radiology reports could include:

- Task and Role: Medical Report Simplifier for simplifying radiology reports.
- Instructions: Please rewrite the following technical radiology report in a simplified version that can be easily understood by a non-specialist audience. Focus on clarity and brevity while retaining all critical medical information.
- Output Format: A simplified, concise, and non-technical summary of the report.
- Cautionary Note: The target audience does not have a medical background, so avoid medical terms.
- Input: [Insert full text of the radiology report]

In contrast, the prompt for a knowledge retrieval-based prompt for the differential diagnosis using image descriptions would look like this:

- Task and Role: As a Diagnostic Support Assistant for finding the correct differential diagnosis.
- Knowledge Retrieval: [Retrieved text from the database of case reports closely matching the inputted description using zero-shot learning].

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

▶ **Table 1** Examples of responses from ChatGPT using GPT4 to different prompting strategies.

Patient: 65-year-old male
Medical history: Hypertension, atrial fibrillation, type 2 diabetes, ex-smoker. Recent coronary angiography (3 days ago)
Current medications: Aspirin, Atorvastatin, Metformin
Complaint: The patient, a 58-year-old male, presents at the emergency department with a sudden and severe onset of abdominal pain. He describes the pain as intense, crampy, and diffuse across the entire abdominal region. The pain began approximately four hours ago and has progressively worsened. He reports no associated symptoms such as nausea, vomiting, or changes in bowel habits.
Physical examination: The patient appears pale, diaphoretic, and in distress. Examination reveals diffuse tenderness, guarding, and rebound tenderness across the entire abdomen. Bowel sounds are absent. BP: 95/60 mmHg. Pulse: 110 bpm, irregular. Abdomen: absent bowel sounds.
Laboratory investigation: Hb 9.8 g/dl; Hct 30 %; lactate 4.2 mmol/L
ECG : Atrial fibrillation with rapid ventricular response.

Method	Prompt	Answer
Basic prompt	What to do first?	The patient's presentation suggests acute abdomen, possibly due to an ischemic event like mesenteric ischemia. The first step should be to stabilize the patient, including intravenous fluid resuscitation, oxygen supplementation, and rate control for atrial fibrillation. Urgent surgical consultation and abdominal imaging (like a CT scan) are also needed.
Few-shot learning	Case file: {Case} What imaging is needed? Examples to solve the task: Case file: Incidentally detected pancreatic cyst. Imaging: MRI of abdomen without and with IV contrast with MRCP Case file: Severe twisting trauma to the knee. Initial imaging. Imaging: Radiography of the knee	CT of the abdomen with IV contrast.
Precision prompt, no further knowledge added to LLM	Case file: {Case} Question: Is imaging usually appropriate for this and if so, specify only which imaging modality and body region is most appropriate and if contrast is needed and if so, which contrast phases.	Yes, imaging is appropriate in this case. A CT scan of the abdomen and pelvis with intravenous contrast is most appropriate. Both non-contrast and contrast-enhanced phases (arterial and venous) are needed to properly visualize potential abnormalities such as ischemia or infarction in the intestines.
Precision prompt and zero-shot learning (manually)	We have provided context information below: {Manually insert text from the ACR mesenteric ischemia guideline} Given this information, please answer the following question concisely and clearly and make a definitive statement. {Precision prompt + case file}	Yes, imaging is usually appropriate for this case. The most appropriate imaging modality would be CT angiography (CTA) of the abdomen and pelvis. This should be done with intravenous contrast, including both arterial and portal venous phases.
Precision prompt and zero-shot learning (automatic knowledge retrieval using vector index)	We have provided context information below: {Automatically retrieved context from guidelines using cosine similarity analysis based on text embedding} Given this information, please answer the following question concisely and clearly and make a definitive statement. {Precision prompt + case file}	Yes, imaging is usually appropriate for this patient. The most appropriate imaging modality would be a computed tomography (CT) scan of the abdomen and pelvis. Given the patient's presentation and elevated lactate levels, there is a suspicion of mesenteric ischemia. Therefore, a dual-phase (arterial and portal venous phase) CT scan with intravenous contrast is recommended to adequately evaluate the mesenteric vasculature. References: Acute nonlocalized abdominal pain (page 10, 15, 16, 3, 4, 6) Imaging of mesenteric ischemia (page 3, 8) These sources can be linked to retrieve the used context or provide a link to enable the user to directly jump to the used source and page e. g. Page "10" would link to https://acsearch.acr.org/docs/69467/Narrative/#page=10.

- Instructions: Using the retrieved knowledge mentioned above, provide the most likely differential diagnoses for the following image descriptions. Consider common and rare conditions and justify each diagnosis with relevant findings.
- Output Format: List of 3 differential diagnoses each with justifications based on the image descriptions provided.
- Cautionary Note: Do not state differential diagnoses with low probability

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Input: [Insert detailed descriptions of radiological images]

Challenges related to data privacy, potential biases, and ensuring consistent accuracy could hamper broad usage of LLMs in the medical domain [17, 18]. This might be addressed by the recent prompt engineering advances ranging from precision prompts to zero-shot learning. Enhanced with vector store retrieval, the remarkable adaptability and potential of LLMs in the rapidly evolving field of medicine become evident [7, 8, 16]. With these advances, the use of prompt engineering and few/zero-shot learning techniques can address these challenges. Due to the improved performance of LLMs, smaller locally deployed networks that do not rely on cloud computing may become feasible. This reduces privacy concerns since sensitive information can be processed locally. Particularly with zero-shot learning approaches, response bias can be mitigated by careful selection of data sources under control of local needs. Templates added by precision prompting or few-shot learning enable a consistent response structure, and robust knowledge retrieval in zero-shot learning improves the consistency of response content. In addition, revealing the source of information on which the answer is based by presenting metadata and hyperlinks to the data has the potential to overcome the current lack of trustworthiness of LLM results. LLMs usually do not admit a lack of knowledge, which in turn leads to so-called hallucinations [19]. Zero-shot approaches thus allow for a deeper insight into the validity of the answer and decision-making. Moreover, the application of LLMs in clinical practice would clearly benefit from the open sourcing of data and model source code to ensure reproducibility, transparency, and independence.

In summary, optimized prompting strategies leverage the potential of LLMs and facilitate their application for medical tasks. Therefore, basic knowledge on prompt engineering is the foundation for creating more accurate and tailored responses. Improvements are gained by moving from precision prompts to advanced techniques like few-shot and zero-shot learning. Emerging embedding-based retrieval mechanisms further amplify the LLMs' capabilities, making them invaluable tools in the medical field.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] OpenAI Platform [Internet]. [zitiert 31. August 2023]. Verfügbar unter: https://platform.openai.com

[2] Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2 (2): e0000198

[3] Lyu Q, Tan J, Zapadka ME et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. 18 2023; 6 (1): 9

[4] Amin KS, Davis MA, Doshi R et al. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology 2023; 309 (2): e232561

[5] Schmidt S, Zimmerer A, Cucos T et al. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg [Internet] 2023. doi:10.1007/s00402-023-05113-4

[6] Jeblick K, Schachtner B, Dexl J et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports [Internet]. arXiv 2022. doi:10.48550/arXiv.2212.14882. http://arxiv.org/abs/2212.14882

[7] Sushil M, Kennedy VE, Miao BY et al. Extracting detailed oncologic history and treatment plan from medical oncology notes with large language models [Internet]. arXiv 2023. doi:10.48550/arXiv.2308.03853. http://arxiv.org/abs/2308.03853

[8] Russe MF, Fink A, Ngo H et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep 2023; 13 (1): 14215

[9] Wang J, Shi E, Yu S et al. Prompt Engineering for Healthcare: Methodologies and Applications [Internet]. arXiv 2023. doi:10.48550/arXiv.2304.14670. http://arxiv.org/abs/2304.14670

[10] White J, Fu Q, Hays S et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT [Internet]. arXiv 2023. doi:10.48550/arXiv.2302.11382. http://arxiv.org/abs/2302.11382

[11] Pinto dos Santos D, Brodehl S, Baeßler B et al. Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. Insights into Imaging 2019; 10 (1): 93

[12] Ye S, Hwang H, Yang S et al. In-Context Instruction Learning [Internet]. arXiv 2023. doi:10.48550/arXiv.2302.14691. http://arxiv.org/abs/2302.14691

[13] Brown TB, Mann B, Ryder N et al. Language Models are Few-Shot Learners [Internet]. arXiv 2020. doi:10.48550/arXiv.2005.14165. http://arxiv.org/abs/2005.14165

[14] Liu Z, Yu X, Zhang L et al. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4 [Internet]. arXiv 2023. doi:10.48550/arXiv.2303.11032. http://arxiv.org/abs/2303.11032

[15] Jin Q, Dhingra B, Cohen WW et al. Probing Biomedical Embeddings from Language Models [Internet]. arXiv 2019. doi:10.48550/arXiv.1904.02181. http://arxiv.org/abs/1904.02181

[16] Rau A, Rau S, Zoeller D et al. A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology 2023; 308 (1): e230970

[17] Geis JR, Brady AP, Wu CC et al. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. Radiology 2019; 293 (2): 436–440

[18] Keskinbora KH. Medical ethics considerations on artificial intelligence. Journal of Clinical Neuroscience 2019; 64: 277–282

[19] Goddard J. Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. The American Journal of Medicine 2023; 136 (11): 1059–1060

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.