Article

Not peer-reviewed version

Understanding the Inner Workings of Large Language Models in Medicine

Georg Fuellen *, Hans Jarchow, Johann-Christian Põder

Posted Date: 8 October 2025

doi: 10.20944/preprints202510.0630.v1

Keywords: Natural Language Processing; Large Language Models; Medical Informatics; Medical Ethics; AI Ethics; Clinical Decision-Making

 doi:10.20944/preprints202510.0630.v1

Article

Understanding the Inner Workings of Large Language Models in Medicine

Georg Fuellen [1]*, Hans Jarchow [1] and Johann-Christian Põder [2]

[1] Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany

[2] Ethics in Theology and Medicine, Faculty of Theology, University of Rostock, Rostock, Germany

* Correspondence: fuellen@alum.mit.edu

Abstract

Background: Large language models (LLMs) are increasingly influencing medical practice, education, and research. Their responsible integration into healthcare requires expertise in medical, ethical, practical, and theoretical domains. **Objectives:** This article examines how theoretical knowledge of LLMs and their internal mechanisms enhances the interpretation of model outputs in medical contexts. **Methods**: We prompted GPT-o1 to generate examples illustrating how understanding transformer architecture can facilitate output interpretation. Key topics were extracted from its responses, and illustrative cases were validated using Consensus.app, an AI-based web-search tool. **Results**: Five key topics were identified: (1) anticipating contextual focus in medical reasoning, (2) explaining "generic" or "textbook" responses, (3) understanding strengths and weaknesses in differential diagnosis, (4) explaining ambiguous or contradictory responses, and (5) identifying hallucinations in unfamiliar scenarios. Case examples highlight both benefits and limitations, including accurate attention to salient clinical details, reliance on generalized patterns, risks of base rate neglect in differential diagnosis, challenges of ambiguous prompts, and hallucinations in rare or underrepresented cases. **Conclusions**: A theoretical understanding of LLMs is crucial for responsible clinical integration. Distinguishing between well-represented (short head) and underrepresented (long tail) knowledge, recognizing generic responses, and identifying hallucinations are essential competencies. Coupled with medical and ethical expertise, these skills will enable healthcare professionals to leverage LLMs effectively while mitigating risks.

Keywords: Natural Language Processing; Large Language Models; Medical Informatics; Medical Ethics; AI Ethics; Clinical Decision-Making

Introduction

The transformative potential of generative AI and, specifically, of large language models (LLMs) is reshaping contemporary medical theory, practice and education. It profoundly influences the future of healthcare as a system, and of its stakeholders. Alongside recognizing the strengths of AI, important questions about its accuracy, and about ethical implications in real-world clinical use have emerged. The medical profession must shape its adoption, despite substantial uncertainties about future opportunities and risks[1].

As LLM-generated content continues to improve, there is a growing risk that decreasing error rates could make us less cautious in vetting its outputs. This stands in contrast to evidence showing that, in certain tasks, LLMs may already deliver best results without human intervention[2]. To harness the advantages of LLMs sensibly, ensuring a responsible integration of AI in healthcare, we must strive for profound understanding, and it is thus crucial to develop and maintain expertise and skills in four key areas:

1. Medical expertise: To be able to critically evaluate the validity of LLM-generated output.

 doi:10.20944/preprints202510.0630.v1

2. Ethics expertise: To identify potential risks and to address instances of ethical dilemmas and violations of medical-ethical norms.
3. Practical knowledge: Familiarity with LLMs, by informed and critical use.
4. Theoretical knowledge: Knowledge of how LLMs operate, which allows for a more nuanced evaluation of LLM-generated content.

The focus here is on the fourth skill, understanding the inner workings of LLMs to better interpret the contents they generate. An example is to discern whether the generated content reflects commonly available and well-represented information in their training data (short head knowledge), or whether it refers to underrepresented knowledge (long tail knowledge), potentially causing erroneous responses. These issues are at best mentioned in passing in recent reviews[3]. We set out to find examples of such benefits, employing LLMs for this task – in a self-referential fashion – and we then scrutinized the key topics and examples they provided. Furthermore, we will identify situations where such benefits are the most likely. A specific benefit from better understanding the inner workings of generative AI – notwithstanding our limitations of understanding, right now and even more so in the future – is that it allows more effective benchmarking by enhancing human oversight, not just flagging potential errors but giving specific explanations for these[1].

Objectives

To explore how theoretical knowledge of LLMs, specifically their architecture and mechanisms, can contribute to better interpretation of generated content in medical contexts, and to identify examples and situations where this understanding offers the most benefit.

Methods

Theoretical and Practical Framework

LLMs are neural-network-based text processing tools composed of multi-head self-attention layers and multi-layer perceptrons (MLPs), trained on vast amounts of input text, aiming to predict the next word. During text processing, the multi-head attention mechanism allows the models to maintain more than one focus, each one on specific parts of the text, where each focus may in turn be connecting dispersed fragments of text. The MLPs combine and process the results of the attention layers further. At the end of a processing pipeline of attention layers and MLPs, the next word is predicted.

As of mid-December 2024, multiple LLMs existed, but OpenAI's GPT-o1 in "professional mode" had – arguably – the most impressive capabilities. Yet, it lacked web-search features, so its references were often hallucinated. To address this issue, we (1) prompted GPT-o1 for examples that demonstrate how understanding the transformer architecture can facilitate the interpretation of model-generated outputs in a medical context, (2) extracted a list of key topics from the headings provided by GPT-o1, and (3) generated well-referenced examples using a separate AI-based web-search engine, that is, Consensus.app.

Model Responses

The AI-based tool Consensus.app proved particularly helpful, yielding a robust set of references without a paid subscription. All prompts and unedited responses are available in the Supplementary Material (see **Supplementary Section A** and **Supplementary Section B**).

GPT-o1 suggested the following key topics:

1. Anticipating contextual focus in medical reasoning.
2. Explaining "generic" or "textbook" responses.
3. Understanding strengths and weaknesses in differential diagnosis.
4. Explaining ambiguous or contradictory responses.
5. Identifying hallucinations in unfamiliar scenarios.

 doi:10.20944/preprints202510.0630.v1

On this basis, Consensus.app then found and referenced two studies: one discussing LLMs in medical rehabilitation (using a stroke case)[4], and another describing how LLMs can facilitate clinical interview analysis through AI-driven synthetic data augmentation[5].

Ethical Considerations

This article does not involve research with human participants or animals. No ethical approval was required.

Results

Based on the results of Consensus.app, we investigated the work of Zhang et al.[4] to address key topics (1), (2), and (5) (see Table 1). In their study, the authors employ a stroke case to evaluate the capability of LLMs in generating rehabilitation recommendations and ICF codes. We also investigated Wu et al.[5] , who describe how LLMs can facilitate clinical interview analysis through their "CALLM" framework, a method for AI-driven synthetic data augmentation. This investigation contributed to the analysis of key topics (3) and (4) (see Table 2).

Table 1. Key Topics (1), (2) and (5), described by GPT-o1, exemplified by Consensus.app.

GPT-o1 explanation for the key topic	Consensus.app explanation for the example	Quotes from the paper (Zhang et al.)	Synopsis
(1) Anticipating contextual focus in medical reasoning			
"Transformers use attention mechanisms to "focus" on specific parts of the input sequence. In a medical dialogue, the model might concentrate heavily on a particular symptom, lab value, or part of a patient's history. Understanding this can help clinicians and medical researchers appreciate why a model's conclusion might hinge on early-mentioned or highly salient details rather than a broad synthesis of all information."	"Understanding the architecture of transformer models, particularly the attention mechanism, can help anticipate how these models focus on specific parts of the input text when generating medical reasoning. For instance, in the context of rehabilitation medicine, ChatGPT-4 was able to generate comprehensive medical and rehabilitation prescriptions by focusing on relevant parts of the input case description, demonstrating its ability to prioritize contextually important information."	"[the LLM] also demonstrated the ability to propose specified approaches for each rehabilitation therapy."	LLMs can indeed prioritize specific characteristics of a case or patient, and complex scenarios can be addressed through the parallel operations of multiple attention heads. While this capability offers advantages in focusing on critical aspects, it may also have drawbacks, particularly when LLMs fail to correctly integrate and synthesize the resulting information.
(2) Explaining "generic" or "textbook" responses			
"MLP layers in transformers combine information gathered by attention heads into higher-level abstractions. When operating on medical queries, these layers may rely on well-learned, "standardized" patterns from training data (e.g., common guidelines or textbook phrasing) instead of tailoring responses to unusual clinical nuances. Understanding the	"In the study on rehabilitation medicine, ChatGPT-4 produced broader and more general prescriptions that were consistent with textbook answers, indicating its reliance on learned generic medical knowledge."	"Compared with standard answers, the large language model generated broader and more general prescriptions in terms of medical problems and management plans, rehabilitation problems and management plans, as well as rehabilitation goals."	Referring to knowledge that is well-represented in the training data (short head knowledge) can result in "generic" or "textbook" responses, raising concerns, however, about their adequacy when addressing atypical cases and patients.

 doi:10.20944/preprints202510.0630.v1

MLP's integrating role explains why a model might revert to a generic standard-of-care response even when presented with a complex or unique patient scenario."			
(5) Identifying hallucinations in unfamiliar scenarios			
"Transformers are trained on patterns within a certain data distribution. When confronted with rare conditions, novel treatments, or unusual clinical contexts, the model's learned patterns may not apply. Attention could be misdirected, and the MLP layers might produce "hallucinated" content because they have no solid internal representation for the out-of-distribution input."	"[...] while ChatGPT-4 made an error in the ICF category, it accurately generated ICF codes, highlighting the model's potential to hallucinate in less familiar contexts."	"A thorough review of the standard clinical ICF code assigned by 2 PMR clinicians was then conducted, comparing it with the table produced by the GPT-4 model (Table II). The 3-digit codes generated by the LLM were accurate (...) However, an error was found when reviewing the case record in the body structures category (s730). The patient had had a stroke, and the original impairment should have been classified as affecting the right precentral gyrus (s110.1), as outlined in the case section. Instead, the table displayed the damage as being in "the upper extremity, left hand." "	LLM responses may exhibit hallucinations when referring to "long tail" knowledge that is not well-represented in the training data. This is hypothesized to be the case for the "body structures category". *Then again, the LLM-generated explanations in this table are not necessarily correct either*. A simpler hypothesis regarding the LLM failure is that it did not know or did not consider that the "body structures category" is supposed to refer to the primary site of damage (the brain), not to the secondary site (the hand). Any lack of knowledge regarding the reporting of ICF categories may thus be attributed to insufficient training data regarding this meta-level information.

Table 2. Key Topics (3) and (4), described by GPT-o1, exemplified by Consensus.app.

GPT-o1 explanation for the key topic	Consensus.app explanation for the example	Quotes from the paper (Wu et al.)	Synopsis
(3) Understanding strengths and weaknesses in differential diagnosis			
"Attention layers help identify connections between symptoms and conditions, while MLP layers synthesize these into coherent outputs. Knowing this pipeline is useful when the model suggests a differential diagnosis. If the model posits an unusual condition, it might be because it latched onto a distinctive symptom that strongly correlated with that condition in its training data—even if that condition is clinically improbable."	"The strengths of transformer models in differential diagnosis can be attributed to their ability to synthesize information from diverse sources, while weaknesses may arise from their lack of real-world clinical experience. The CALLM framework, for example, enhances clinical interview analysis by generating synthetic data that can improve diagnostic accuracy, showcasing the model's adaptability in learning from augmented datasets."	"In automated mental health diagnosis, the scarcity and imbalance of clinical data pose considerable challenges for researchers, limiting the effectiveness of machine learning algorithms. To cope with this issue, this paper aims to introduce a novel clinical transcript data augmentation framework by leveraging large language models (CALLM). The framework follows a "patient-doctor role-playing" intuition to generate realistic synthetic data."	A hypothesis about how LLMs handle differential diagnoses is that multi-head attention may be responsible for the matching of patient data to the sets of symptoms known for disease conditions, but this matching may ignore disease prevalence. Synthetic data may mitigate this weakness because researchers can generate examples following a data distribution they have under control, and provide these examples to the LLM.
(4) Explaining ambiguous or contradictory responses			
"When the patient's presentation is ambiguous or the prompts contain conflicting information, attention mechanisms may	"Ambiguities or contradictions in model outputs can often be traced back to the model's training data or the inherent	"Our "Response-Reason" prompting approach guides LLMs in generating highly authentic clinical	Contradictory responses can be attributed to ambiguous input, ambiguity within the training data, or ambiguity in the

 doi:10.20944/preprints202510.0630.v1

distribute focus across multiple, equally plausible interpretations. The MLP layers may fail to resolve these into a single, authoritative answer. Understanding this helps users interpret uncertain or oscillating responses as a reflection of the model's internal struggle with ambiguity rather than mere randomness."	complexity of medical language. The CALLM framework's use of a "Response-Reason" prompt engineering paradigm aims to generate diagnostically valuable transcripts, which can help mitigate such issues by providing clearer reasoning paths in the model's responses."	interview transcripts for mental disorder diagnosis. This augmentation is tailored to enhance the training dataset, facilitating both FSL[Few-Shot-Learning] and, in certain cases, ZSL[Zero-Shot-Learning]." "This technique [...] encouraged it to elucidate the rationale behind the responses, mirroring the profile and characteristics of a simulated patient."	representation of knowledge by the trained model. Specialized prompting techniques may request that the reasoning path of the LLM is made more transparent, enhancing its reasoning capabilities along the way.

Key topic 1: Anticipating contextual focus in medical reasoning

Key topic (1) illustrates how the LLM's attention mechanism enables it to identify and prioritize relevant information within a text, connecting specific parts of the text that may be far away in the input stream. In fact, this "focus on specific parts" may be done multiple times in parallel by multiple attention heads and integrated across layers until the final output is generated. The attention mechanism thus allows the model to focus on critical elements – such as specific patient symptoms or lab values – while ignoring less pertinent data. GPT-o1 underscores that clinicians and medical researchers can benefit from understanding that model attention might attend to highly salient details at the expense of a broader synthesis. This focused approach can be advantageous in ensuring the model's output is closely aligned with key aspects of a query.

In the example by Consensus.app, Zhang et al.[4] demonstrate how this capacity proves valuable in their example from rehabilitation medicine, where ChatGPT-4 generated targeted intervention plans by focussing on the most relevant details of a patient's presentation.

Key topic 2: Explaining "generic" or "textbook" responses

"Generic" or "textbook" responses arise from a model's tendency to draw on widely represented knowledge learned from its training data ("short head knowledge"). GPT-o1 suggests that, when responding to medical queries, the model's MLPs often rely on well-learned patterns, which can lead to standardized procedures being presented even in atypical clinical situations. This is echoed in the Consensus.app findings, indicating ChatGPT-4's propensity to default to generalized medical knowledge.

Key topic 3: Understanding strengths and weaknesses in differential diagnosis

With regard to key topic (3), GPT-o1 suggests that models sometimes include improbable clinical differential diagnoses due to strong correlations detected between input and training data, irrespective of the prevalence of the diagnosed disease; such base rate neglect may or may not be helpful for an accurate diagnosis[6]. Consensus.app mentions a lack of direct clinical experience of the models, which may refer to base rate neglect. In the example of Wu et al.[5], accuracy was supposedly enhanced by data augmentation using synthetic data as part of the CALLM framework, allowing better differential diagnoses based on more balanced data. However, skeptics may argue that true clinical complexity is difficult to replicate through synthetic data, casting doubt on the broader applicability of AI-generated simulations for real-world clinical settings. Thus, there is an evident tension between synthetic data generation and the complexity of capturing clinical "real-world" scenarios. This also underscores the critical importance of robust validation requirements[7], particularly when significant decisions are to be made by an LLM.

Key topic 4: Explaining ambiguous or contradictory responses

In addressing key topic (4), GPT-o1 links ambiguous or contradictory outputs to unclear or insufficiently specific prompts, whereas Consensus.app posits that ambiguity within the model's

 doi:10.20944/preprints202510.0630.v1

training data is a contributing factor. One solution involves carefully crafted prompts designed to elicit the model's reasoning processes, thereby mitigating confusion. Here, the CALLM framework successfully employs a "Response-Reason" prompting strategy.

Key topic 5: Identifying hallucinations in unfamiliar scenarios

By contrast, "unfamiliar scenarios" engage a model's "long tail" knowledge, where there is a heightened risk of hallucination because the queries may diverge significantly from what could be learned from the training set.

A case in point, discussed by Zhang et al.[4], shows ChatGPT-4 successfully generating International Classification of Functioning (ICF) codes for a stroke patient but misreporting the lesion site. Specifically, the model accurately identified the motor dysfunction in the left hand but failed to report the lesion in the right precentral gyrus. Although the model recognizes that the patient's motor function is impaired, it does not appear to understand that this impairment originates from disrupted motor signals in the brain. As a result, the model interprets the limitation as purely motor-related rather than addressing the underlying neurological cause. Alternatively, we suggest that it may miss the meta-knowledge that the lesion site to be reported here shall refer to the underlying primary lesion, not its secondary consequences.

This distinction, however, is crucial: if a clinical decision-support system fails to report the specific neurological lesion, it may overlook critical rehabilitation strategies that are essential for effective patient care. Consequently, interventions might miss addressing the root cause, leading to slower or less effective patient recovery.

These examples highlight the importance of recognizing not only what a model can accomplish but also where gaps in its knowledge or reasoning may lead to clinically relevant inaccuracies.

Discussion

The implementation of AI, particularly LLMs, in healthcare will drive transformative changes in medical practice and theory while presenting significant challenges. A thorough understanding of the key ingredients of LLMs, based on their underlying architecture, including attention mechanisms and MLPs, can be particularly useful in situations where the model's outputs are unexpected, ambiguous, or counterintuitive, necessitating critical analysis.

This knowledge is also important for addressing ethical concerns in sensitive applications, for example by providing a deeper understanding of where harmful and discriminatory biases can arise. For example, the ability to distinguish whether generated content reflects well-represented (short head) or underrepresented (long tail) information can be critical in identifying potential biases in AI outputs, as most training data reflect majority rather than minority or marginalised perspectives.

This is particularly evident as far as generic responses are concerned (key topic 2), but also intersects with other key topics, such as difficulties in dealing with unfamiliar scenarios (key topic 5). Theoretical knowledge of how LLMs operate can thus contribute significantly to the ethically sound and responsible integration of AI in healthcare.

The existing tension between opportunities and limitations of using LLMs underscores the need for careful evaluation. Given the rapid development of LLMs, acquiring expertise and skills that bridge medical knowledge with AI proficiency and ethics will become increasingly important. Such competencies will enable healthcare professionals to leverage and benchmark the benefits of AI and LLMs while remaining vigilant regarding their limitations[1], and we hope that we can even keep up some understanding also of any advanced AI still on the horizon.

Conclusions

The examples analyzed in this study demonstrate that LLMs hold transformative potential in healthcare, and theoretical knowledge of their architecture and mechanisms is important for interpreting their outputs responsibly. Understanding the balance between short head and long tail

 doi:10.20944/preprints202510.0630.v1

knowledge, recognizing generic responses, and identifying hallucinations are crucial skills. These competencies will empower healthcare professionals to integrate AI effectively while mitigating risks and ensuring ethical standards.

Conflict of Interest: non declared

References

1. Klang E, Tessler I, Freeman R, Sorin V, Nadkarni GN. If Machines Exceed Us: Health Care at an Inflection Point. NEJM AI. 2024;1(10):AIP2400559. doi:10.1056/AIp2400559
2. Ranji SR. Large Language Models—Misdiagnosing Diagnostic Excellence? JAMA Network Open. 2024;7(10):e2440901-e2440901. doi:10.1001/jamanetworkopen.2024.40901
3. McCoy LG, Ci Ng FY, Sauer CM, et al. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review. BMC Medical Education. 2024/10/07 2024;24(1):1096. doi:10.1186/s12909-024-06048-z
4. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. Journal of Rehabilitation Medicine. 09/11 2023;55:jrm13373. doi:10.2340/jrm.v55.13373
5. Wu Y, Mao K, Zhang Y, Chen J. CALLM: Enhancing Clinical Interview Analysis Through Data Augmentation With Large Language Models. IEEE Journal of Biomedical and Health Informatics. 2024;28(12):7531-7542. doi:10.1109/JBHI.2024.3435085
6. Hamm RM. Physicians neglect base rates, and it matters. Behavioral and Brain Sciences. 1996;19(1):25-26. doi:10.1017/S0140525X00041261
7. Fuellen G, Kulaga A, Lobentanzer S, et al. Validation requirements for AI-based intervention-evaluation in aging and longevity research and practice. Ageing Res Rev. Feb 2025;104:102617. doi:10.1016/j.arr.2024.102617