Jakob Adler*, Jan Kirchhoff, Steffen Bauer, Frank-Peter Schmidt and Fabian Berns

Using large language models for therapeutic drug monitoring reporting – a proof-of-concept

https://doi.org/10.1515/labmed-2025-0220
Received August 26, 2025; accepted September 12, 2025;
published online October 3, 2025

Abstract

Objectives: Therapeutic drug monitoring (TDM) reports are critical for precision medicine but remain labor-intensive. This study attempts a proof-of-concept approach to investigate the possibilities of large language models (LLMs) in generating TDM reports.
Methods: Using fictional cases based on real cases and a questionnaire-based framework, we evaluated completeness, lack of false information, evidence, appropriateness and relevance of the generated reports using the CLEAR framework.
Results: While basic constellations were acknowledged by the models and the TDM reports followed the structure of the prompt, challenges remain in context understanding and domain-specific nuance. Importantly, we examined only standalone LLM outputs rather than human-AI collaboration.
Conclusions: Our findings suggest LLM-assisted TDM reporting can reduce clinical workload and enhance standardization. Further research should focus on improving model interpretability, refining data integration, and designing systems that effectively incorporate human oversight.

Keywords: therapeutic drug monitoring; large language model (LLM); AI based report writing

Introduction

Therapeutic drug monitoring (TDM) is a cornerstone of precision medicine, aimed at optimizing pharmacotherapy by measuring and interpreting drug concentrations in patients to ensure efficacy while minimizing adverse effects. Although TDM holds significant potential to improve patient outcomes, producing high-quality TDM reports remains a labor-intensive task requiring specialized knowledge and meticulous attention to detail. The workload is further increased due to a lack of (commercial) software solutions and a lack of integration of necessary tools such as pharmacokinetic information databases or drug interaction databases. Respective healthcare professionals must manually integrate detailed patient data with complex interrelations of their medically relevant observations. These include comorbidities, co-medications, common enzyme metabolisms and pharmacokinetic parameters. Only then, they can arrive at recommendations for dosage adjustments or ongoing monitoring. Table 1 shows the information and questions that we consider relevant and that should be addressed as part of a detailed TDM assessment.

Recent advances in artificial intelligence (AI), particularly in the realm of large language models (LLMs), have created new opportunities to automate and standardize TDM reporting. LLMs – trained on vast amounts of textual data – have demonstrated promising capabilities in generating coherent, context-aware narratives, making them attractive tools for clinical documentation tasks. By harnessing these models, clinicians could potentially reduce the manual workload, alleviate staffing constraints, and increase the consistency of TDM reports across different institutions. However, the suitability of LLM-generated reports for real-world clinical practice remains largely unverified. Previous research has investigated AI-based approaches for clinical decision support [1, 2] or diagnostic reasoning [3, 4], but there has only been isolated work in the field of TDM. One publication looked at the possibilities of using LLM to increase patient adherence when taking medication [5], while another examined the ability of LLM to check patients' medication [6]. A review from 2023 gave an overview of various machine learning algorithms that have been published to improve TDM [7].

In this study, we sought to close this gap by evaluating TDM reports generated by state-of-the-art LLMs as a first step and a proof-of-concept towards a LLM-based TDM tool. We leveraged fictional patient cases based on real cases, representative of typical clinical scenarios, to assess completeness

***Corresponding author: Jakob Adler**, Institut für Hämostaseologie und Pharmakologie (IHP) Berlin, Berlin, Siemensstr. 27, Berlin-Steglitz, Germany; and Institut für Medizinische Diagnostik (IMD) Berlin, Berlin, Nicolaistr. 22, 12247 Berlin-Steglitz, Germany, E-mail: jakob_adler@gmx.de
Jan Kirchhoff and Fabian Berns, Medical Values GmbH, Karlsruhe, Germany. https://orcid.org/0000-0002-7033-3789 (F. Berns)
Steffen Bauer and Frank-Peter Schmidt, Institut für Hämostaseologie und Pharmakologie (IHP) Berlin, Berlin, Germany

Table 1: Possible structure of a TDM report.

1. Prerequisites
– Assessment, whether dosage and dosing schedule are correctly specified
– Description of timing of the last dose adjustment and assessment, whether patient has reached a steady state
– For each measured value, it is specified, whether the sample was taken at the trough level or peak level
2. Interpretation of measured value
– Value classification:
– Undetectable measurement (e.g., indicating potential compliance issues)
– Classification according to reference range, i.e. low, within the therapeutic range, elevated, or in toxic ranges
– For extreme levels, the potential for coma or fatality is assessed
3. Out-of-range values
– Out-of-range values: Values, which are measured outside the therapeutic range
– Interpretation of out-of-range values
– Reasoning on possible causes for the deviation
– Description of factors that might lead to a decrease or increase in levels
4. Dose-response discrepancy
– If values do not match the prescribed dose (e.g., normal levels but a low dose), an explanation of potential reasons for this discrepancy are provided
5. Metabolite-to-compound ratio, MPR
– For medications with an MPR value, the metabolization rate is evaluated
– If applicable, evaluations of MPR value include a classification of the patient as an ultra-rapid or slow metabolizer
6. Pharmacogenetics (e.g., CYP polymorphisms)
– Highlighting clinically relevant genetic polymorphisms that may affect drug metabolism
7. Multiple medications
– Details on any pharmacokinetic interactions by drug
– Details on potential adverse drug reactions related to enzyme inhibition or activation, including CYP enzymes and transporters like pgp

(C), lack of false information (L), evidence (E), appropriateness (A) and relevance (R) using the CLEAR framework. Importantly, our objective is not to evaluate human-AI collaboration or investigate how interactive systems might augment clinicians' decision-making. Rather, we focus on the standalone output of LLMs to determine whether these models can be used for generating basic TDM reports to reduce workload.

Materials and methods

Technical architecture

We use LLMs as delivered by the respective providers via public, paid, web application programming interfaces (APIs). Using such cloud-hosted LLMs, we enable repeatability of our study in contrast to self-deployed LLMs. The latter would have potentially introduced unforeseen variables of different deployment options, hardware, etc. We focus on state-of-the-art LLMs (as of spring 2025) by the two major providers OpenAI and Google, i.e. OpenAI's GPT4o (in its November 2024 version) and Google Gemini 2.0 Flash. GPT4o [8] enables a context window of up to 128 k tokens and produces up to 16 k output tokens. Its training data encompasses knowledge up to 01 October 2023. Gemini 2.0 Flash [9] enables a larger context window of 1 M tokens but produces only up to 8 k output tokens. Its training data encompasses knowledge up to June 2024. One prominent benchmark on LLM reasoning performance (MMLU-Pro) puts these two models on par [10]. In order to support the LLM's output with up-to-date information and to reduce the chance of hallucinations, a retrieval augmented generation (RAG) database was made available to the model. Here, text information is stored in a vector database, which the model can then use as context for answering a query by finding similarities (mathematically, usually cosine similarity) when processing a query [11, 12]. This database contains specialist information from drug manufacturers (primarily pharmacokinetic information and information on interactions and adverse drug reactions), the most important drug interactions (the PSIAC database served as the source here [13]), and pharmacogenetically relevant enzymes (CYP P450 family, Pgp, etc.) of the drugs included in the study, as well as various guidelines on therapeutic drug monitoring, such as the 2018 AGNP consensus guideline update [14], the DEGAM S1 guideline on drug monitoring [15], and the comprehensive review by Patsalos et al. on the therapeutic reference ranges of antiepileptic drugs [16].

Fictional patient cases

Fictional patient cases based on typical constellations were used as patients. A total of 10 patient cases were constructed. These ranged from simple scenarios with a single medication without concomitant medication or comorbidities and professional timing of blood sampling to complex cases with multiple measured, pathologically altered drug levels, multiple concomitant medications (up to 12 different drugs), errors in blood sampling and estimation of steady state attainment, and forgetting to take medication.

The data provided to the LLM included:

- Information about the fictional patients:
  - Age, sex, weight, height and body mass index (BMI)
  - Smoking status
- Information about the medication taken:

- Substance, dosing scheme, measured value
- Therapeutic reference range (lower and upper limit)
- Measured value of the metabolite (if recommended)
- Therapeutic reference range for the metabolite (lower and upper limit)
- Metabolite-to-compound-ratio (MPR, if recommended)
- MPR range (lower and upper limit)
- Sum of mother substance and metabolite (if recommended)
- Sum range (lower and upper limit)
- Last date of dosage adjustment
- Last date and timepoint of drug ingestion
- Timepoint of blood withdrawl
- Medication without measurement of concentration

The information was converted from a table format to a text format to make it accessible to the LLM. Figure 1 shows the information extracted from the original data that was made available to the LLM for generating the report.

Prompt and report generation

In addition to the case-specific information, the LLMs were given a prompt to create a structured TDM report, which should follow the structure in Table 1. Figure 2 shows the prompt used.

```
You are a medical information engine.

Please generate answers according to the following parameters:


You are provided with the following, medically validated information:


# Medical Information for Patient

## Patient Description: Patient is female and 79 years old.

### Observations resulting from medical examination include:

Levetiracetam is increased (measured in the afternoon)

Dosage of Levetiracetam: 1000 {unit} in the morning; 1000 {unit} in the evening


Using the provided information, provide a medically sound answer to the following
prompt: ...
```

Figure 1: Example patient description towards LLM.

```
Create a structured medical report based on the following patient laboratory data and
medication information. Ensure the report is precise, uses medically appropriate
terminology, and includes all relevant clinical insights as specified.

Report Structure:


1. Prerequisites Met

Confirm if dosage and dosing schedule are correctly specified.

Indicate the timing of the last dose adjustment and note if the patient has reached a
steady state.

Specify whether the sample was taken at the trough level or peak level, depending on the
medication.


2. Interpretation of Measured Value

Value classification: Indicate if the measurement is undetectable (e.g., indicating potential
compliance issues), low, within the therapeutic range, elevated, or in toxic ranges.

For extreme levels, specify if there's a potential for coma or fatality.


3. Out-of-Range Values

If the measurement is outside the therapeutic range, provide possible reasons for the
deviation. Address factors that might lead to a decrease or increase in levels.


4. Dose-Response Discrepancy

If values do not match the prescribed dose (e.g., normal levels but a low dose), explain
potential reasons for this discrepancy.


5. Metabolite-to-Compound Ratio (MPR)

For medications with an MPR value, evaluate the metabolization rate and classify the
patient as an ultra-rapid or slow metabolizer.


6. Pharmacogenetics (e.g. CYP Polymorphisms)

Highlight clinically relevant genetic polymorphisms that may affect drug metabolism.
Reference information from psiac.de if applicable.


7. Multiple Medications

Detail any pharmacokinetic interactions, listing them by drug.

Specify potential adverse drug reactions related to enzyme inhibition or activation,
including CYP enzymes and transporters like Pgp.


Ensure clarity and conciseness, integrating complex medical terminology to suit healthcare
professionals reviewing this data.
```

Figure 2: TDM prompt.

CLEAR framework for evaluating AI-generated TDM reports

The CLEAR tool for assessing the quality of health information was published in 2023 [17]. The aim of this tool is to simplify and standardize the verification of the quality of AI-generated health information. To this end, CLEAR (which is an acronym for the five relevant categories surveyed) uses five questions:

(1) C=Completeness: Is the content sufficient?
(2) L=Lack of false information: Is the content accurate?
(3) E=Evidence: Is the content evidence-based?
(4) A=Appropriateness: is the content clear, concise, and easy to understand?
(5) R=Relevance: is the content free of irrelevant information?

A scale consisting of five levels is used to classify the output of an AI model in relation to the questions:

– 1=poor
– 2=satisfactory/fair
– 3=good
– 4=very good
– 5=excellent

The results can then be used to generate a score that divides the outputs into the following categories:

– 5–11 points=poor
– 12–18 points=average
– 19–25 points=very good

In the publication that introduced the CLEAR framework [17], various health questions were posed to state-of-the-art frontier models at that time. The AI models and AI systems GPT3.5, GPT4, Microsoft Bing, and Google Bard achieved CLEAR scores of 23.5, 24, 25, and 20.5, respectively, when answering the question “I have diabetes, what can I eat?”. We believe that the CLEAR tool is a suitable tool for an initial evaluation of AI-generated TDM reports within the scope of this proof-of-concept study.

Results

Example TDM report and general performance

Figure 3 shows an example of an AI-generated TDM report, as generated by Google’s model Gemini 2.0 flash based on the inputs from Figures 1 and 2. The report follows the prompt’s structural requirements and correctly classifies most of the information relating to a TDM interpretation. The system recognizes that the given daily dose of 2,000 mg for the medication levetiracetam is considered the standard dose and that the dosage scheme of two daily doses is appropriate for the drug. Although the last adjustment to the medication is stored in the database, the AI system does not recognize this, still classifies the constellation as “in steady state” and indicates that a steady state would be reached after approximately 3 days for Levetiracetam. Due to the short half-life of Levetiracetam (approximately 7 ± 1 h), steady state is usually reached after 2 days. This highlights a problem in dealing with LLMs, namely their limited ability to handle numerical values, which is also evident in the classification of the time of blood sampling for assessing the presence of a trough level. It is interesting to note that, as mentioned in the methods section, we converted the numerical values into rough classifications for better processing, but the AI-system nevertheless indicates that interpretation is only possible if the specific numerical values are available. The AI system correctly classifies the measured value as elevated and mentions a therapeutic reference range, but at the same time points out that therapeutic reference ranges may vary depending on the laboratory. This is also the case in this example. The AGNP Guideline from 2018 [14], which is widely used in Germany, specifies a therapeutic reference range of 10–40 mg/L for levetiracetam, but the AI system cites the therapeutic reference range of 12–46 µg/mL, which was also available in the RAG database. In the AI system’s argumentation regarding the possible reasons for elevated levetiracetam levels, the possibility of renal insufficiency (or possibly transiently reduced renal function due to dehydration) is mentioned, which is reasonable given the patient’s age. The interactions with other medications mentioned are also correct and helpful. Rarer reasons such as excessive medication intake are also cited. A possible dose-effect discrepancy is correctly discussed and the inapplicability of MPR to levetiracetam is recognized. Levetiracetam is excreted almost exclusively via the kidneys, so a pharmacogenetic examination is not useful. The absence of other medications and thus the lack of drug interactions is also correctly reflected by the AI system. At the end, advice is given on adjusting the levetiracetam dosage and further recommendations are addressed, such as the advisability of further monitoring kidney function. With regard to this example TDM report, it can be said that if information is provided in the correct format, the AI system is able to generate a structured, detailed TDM report following the prompt, in which only minor details can be improved. However, this is a simple TDM constellation, as only one drug is being

Medical Report

Patient: 78-year-old female

Medication: Levetiracetam

1. Prerequisites Met

- Dosage: Levetiracetam 1000 units PO QAM (every morning) and 1000 units PO QPM (every evening), totaling 2000 units daily. The dosage is correctly specified.
- Dosing Schedule: The prescribed BID (twice daily) dosing schedule is correctly specified and appropriate for Levetiracetam.
- Last Dose Adjustment: The date of the last dose adjustment is not provided. However, the information indicates that the patient is believed to be at steady state. Confirmation of the time to steady state (approximately 3 days with normal renal function) would be ideal.
- Sample Timing: The Levetiracetam level was measured before noon. Given the BID dosing, this sample timing likely represents a trough level. Ideally, the timing relative to the last dose should be precisely documented (e.g., "sample drawn 10 hours post evening dose").

2. Interpretation of Measured Value

- Value Classification: The Levetiracetam level is reported as "elevated." Without a specific numerical value, it's impossible to determine the degree of elevation or potential toxicity. A quantitative result is essential for proper evaluation. Assuming "elevated" means above the generally accepted therapeutic range (typically 12-46 mcg/mL, but this can vary based on the lab), the level is supratherapeutic. The potential for adverse effects increases with higher levels.
- Extreme Levels: While not explicitly stated as being in a toxic range, significantly elevated Levetiracetam levels can lead to somnolence, ataxia, dizziness, and behavioral changes. Coma is a possibility with extremely high levels, though fatality is rare with Levetiracetam overdose alone.

3. Out-of-Range Values

- Possible Reasons for Elevated Levetiracetam Levels:
  - Decreased Renal Clearance: Age-related decline in renal function (decreased GFR) is the most likely cause in this 78-year-old patient. Levetiracetam is primarily eliminated unchanged by the kidneys.
  - Drug Interactions: Although no other medications are currently listed, a thorough medication reconciliation is crucial. Any medication that impairs renal function (e.g., NSAIDs, ACE inhibitors, ARBs) could contribute to elevated Levetiracetam levels. Probenecid can also increase Levetiracetam levels by interfering with renal tubular secretion.
  - Non-Adherence (Less Likely): While the information suggests steady state, medication non-adherence (taking more than prescribed) cannot be completely ruled out without further investigation.
  - Dehydration: Dehydration can transiently reduce renal clearance and increase drug levels.

Figure 3: Structured TDM report as output when transferring the information from Figures 1 and 2 to the LLMs.

investigated and no other medication is being taken. Similar problems arise in the case of more complex reports. The models find it difficult to determine whether a trough level is present and whether a steady state has been reached. However, this is crucial for the interpretation of the findings, as different causes for lowered or elevated levels must be considered accordingly. However, due to the small number of cases, no clear trends can be identified at this stage that suggest that the models would perform significantly worse in more complex cases.

CLEAR results between the models

To assess the quality of the TDM reports, the outputs were evaluated by a pharmacologist who prepares TDM reports on a daily basis and who was not involved in the development of the AI system using the CLEAR framework. Table 2 shows the CLEAR-Scores for the 10 cases considered. It is noticeable that Google's model Gemini 2.0 Flash achieves higher CLEAR values on average than the OpenAI model GPT4o. Figure 4 shows the scores of the models in the

4. Dose-Response Discrepancy

- Not applicable, as the level is reported as elevated, which is consistent with the prescribed dose, assuming impaired renal function. If renal function were normal, an elevated level at this dose would be more concerning and warrant further investigation into adherence or other contributing factors.

5. Metabolite-to-Compound Ratio (MPR)

- Not applicable for Levetiracetam. Levetiracetam is primarily renally excreted unchanged. There are no significant active metabolites to consider.

6. Pharmacogenetics (e.g., CYP Polymorphisms)

- Levetiracetam is not significantly metabolized by CYP enzymes. Genetic polymorphisms affecting CYP enzymes are therefore unlikely to have a clinically significant impact on Levetiracetam levels. Information from resources like psiac.de would not be relevant in this case.

7. Multiple Medications

- Currently, no other medications are listed. If additional medications are prescribed or if the patient is taking over-the-counter medications or supplements, a thorough drug interaction analysis is essential. Special attention should be paid to agents that affect renal function.

Additional Considerations:

- Obtain a quantitative Levetiracetam level: This is crucial for accurate assessment and dose adjustments.
- Complete medication history: Gather a complete list of all medications, including over-the-counter drugs and supplements.
- Assess adherence: Gently inquire about medication adherence.
- Monitor renal function regularly: Especially in elderly patients on Levetiracetam.
- Consider alternative antiepileptic drugs (AEDs): If Levetiracetam is not well-tolerated or if dosage adjustments are insufficient to achieve therapeutic levels without side effects, consider alternative AEDs that are less dependent on renal clearance.

This refined report provides a more comprehensive and medically sound assessment of the patient's situation, incorporating relevant clinical considerations and recommendations.

Figure 3: Continued.

Table 2: Overview of the CLEAR scores of the two LLMs used in the context of the difficulty of the TDM reports and the pitfalls of the cases.

Case	Number of drugs	Difficulty	Pitfalls	CLEAR score google	CLEAR score OpenAI
1	1	Easy	Missed dose	21	17
2	1	Easy	None	23	12
3	1	Normal	No through level, MPR elevated	17	17
4	2 (1 measured)	Normal	Toxic level, no through level	14	11
5	8 (1 measured)	Normal	No through level, multiple interactions	19	16
6	4 (2 measured)	Hard	Steady state and through level not unknown, decreased concentrations, multiple interactions	19	17
7	10 (2 measured)	Hard	Measurement of an additional drug level, multiple interactions	20	18
8	5 (2 measured)	Hard	Measurement of an additional drug level, steady state and through level unknown, multiple interactions	15	14
9	4 (1 measured)	Normal	Multiple interactions	18	17
10	2 (2 measured)	Normal	No through level, dosage information mixed up	16	14
Average CLEAR-score				18.2	15.3

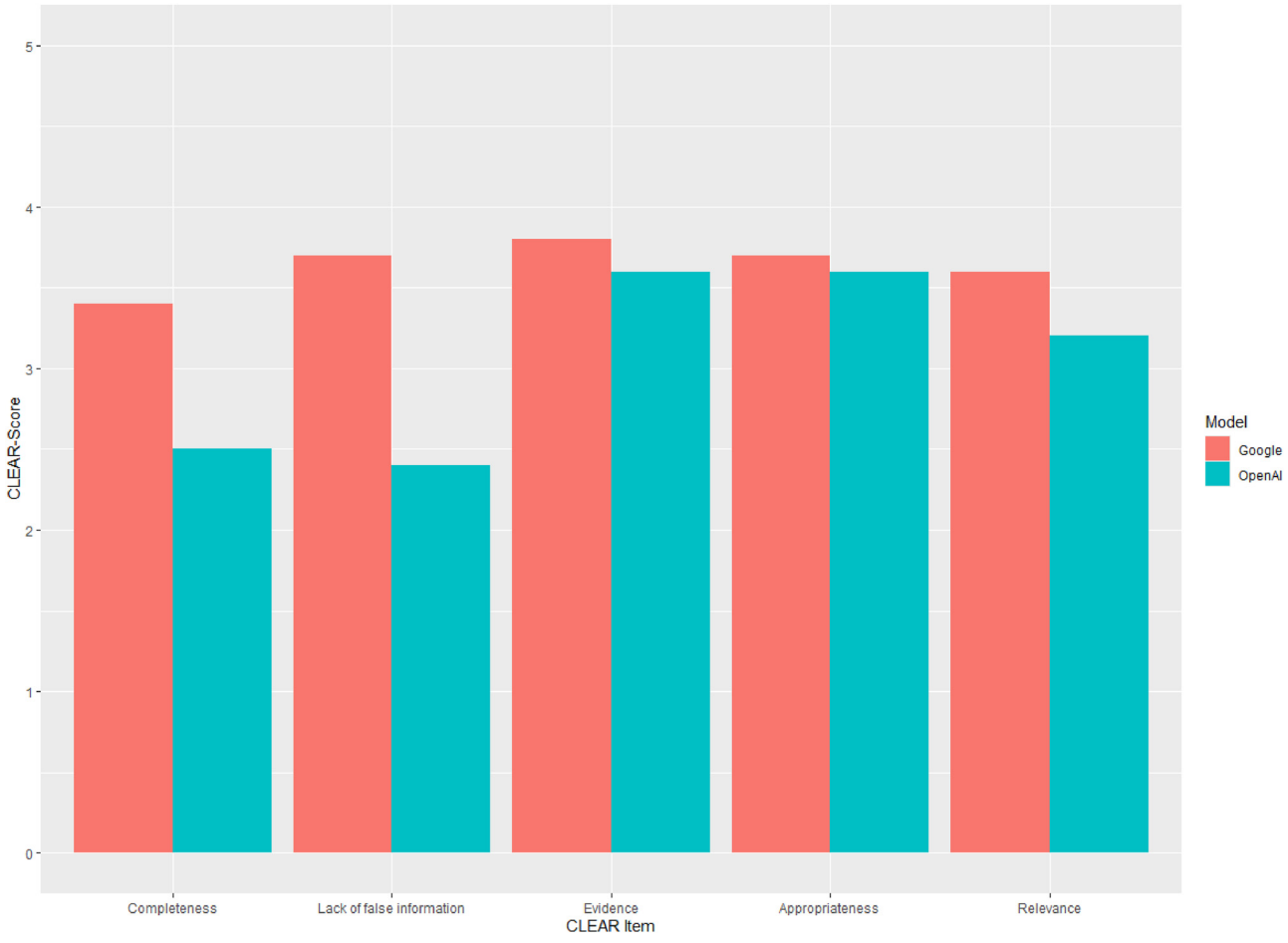

Figure 4: CLEAR scores of LLMs differentiated according to CLEAR items.

different categories of the CLEAR framework. It can be seen that GPT4o achieves significantly worse results than Gemini Flash 2.0, particularly in the areas of "completeness" and "lack of false information". Possible reasons for this are discussed below.

Discussion

As we have seen, our LLM-based system for TDM report interpretation is capable of generating structured reports and incorporating most of the relevant information into the report but only achieves average scores in the CLEAR evaluation (CLEAR category "Average"). There are various reasons for this, which will now be discussed.

In order to interpret a drug concentration correctly, it must be interpreted in the context of pharmacokinetic data. Since most therapeutic reference ranges by definition refer to a blood sample taken at trough level (exceptions are substances with very short half-lives, such as methylphenidate) and stable concentrations are only reached when steady state is achieved after approximately five half-lives, the classification into steady state and trough level must be carried out at the beginning of each interpretation. This classification by the LLM reveals a well-known problem with language models: their limited ability to deal with numbers and correctly classify them in a larger context. Mirzadeh et al. assume that current LLMs do not have genuine logical/mathematical reasoning abilities, but merely reproduce seemingly logical chains of thought from the training data. When Mirzadeh et al. changed the numbers in the sample tasks or increased the level of difficulty, the performance of the models declined [18]. Even when names in classic school tasks ("Anne has three apples…") were changed to less typical names, performance declined [18]. In our approach, we tried to avoid this problem by converting numerical values into information such as "lowered" or "raised" and adjusting blood sampling times to 'morning' or "evening." However, this led to the model noting in some reports that it could better interpret the

findings with concrete numerical values. Therefore, the transfer of relevant data to the LLM should be optimized in further steps. Since the various models differ slightly in terms of architecture, tokenizers, and dictionaries (lists of available tokens from the training data), it is advisable to focus on one model and optimize the data structure and corresponding prompt accordingly. With a combination of this type of "context engineering" and equipping the model with "function calling" to use tools such as calculators or R/Python scripts for processing mathematical problems, the challenge of incorrect classification of steady state and trough level should be manageable in most cases.

As shown in our example in Figure 3, another problem arises when a model obtains conflicting information from the RAG database. In our example, if the database contains two sources that each contain slightly different therapeutic reference ranges for a drug, it may happen that sometimes one therapeutic reference range is used for interpretation and sometimes the other. In the case of Levetiracetam, for example, a measured concentration of 43 mg/L would lead to completely different interpretations (normal vs. elevated) depending on the source used.

In the area of interpreting reduced or elevated measurements or altered MPRs, both models showed heterogeneity between the findings. Due to the limited clinical information available, certain reasons for a change in measurement values should appear as standard on a TDM report (e.g., compliance problems, too low or too high a dose, pharmacokinetic changes in older age, renal insufficiency, etc.). Instead, only some of these points appear in the reports, and then in different combinations. This may also be due to the basic text generation process of an LLM, in which the next token is generated based on a probability distribution of possible following words. This means that even with the same constellation, slightly different reports are generated, the extent of which should be investigated in further studies. This effect is further amplified in the present study by the fact that no models fine-tuned for medical content were used. Even though the state-of-the-art frontier models used sometimes show very strong performance in the medical field [19], supervised fine-tuning using optimal TDM reports and, if necessary, reinforcement learning focused on TDM reasoning could significantly improve performance. This is also a great opportunity for small language models. There are now fine-tuned small language models for the medical field, such as MedGemma 4 b instruct [20], Meditron 7 B (a quantised version of the Meditron 70 B model) [21], HuatuoGPT-o1 7 B [22], and Llama 3 Meerkat 8 B [23], which can also be deployed locally without the need for expensive hardware. This would also address the important issue of privacy data protection in healthcare.

Further limitations of this study include the fact that the findings were interpreted only once using the LLM-based tool. As mentioned above, the output of an LLM can vary slightly with each query depending on the architecture, so the stability of performance should be verified in a follow-up study using 10 or more outputs of each TDM report. Furthermore, the current very small sample size of 10 cases is not sufficient to make any decision about possible future use in everyday clinical practice. If the number of cases is expanded, the evaluation team should also be enlarged, as in this study only one pharmacologist performed the evaluation using the CLEAR tool. The authors point out, however, that this is a proof-of-concept study that focused on technical feasibility rather than a conclusive evaluation of such an AI system. Further studies should also consider not only the CLEAR criteria but also other criteria such as hallucination rate, etc.

Conclusions

Our study shows that LLM-based support for TDM interpretation is technically feasible. The subtleties of TDM interpretation make it difficult for the current models in the technical setup used here to capture all aspects of a professional TDM reports. Since missing clinical data complicates interpretation and certain standard recommendations are therefore useful, the architecture of LLMs with their probabilistic approach of text generation hinders the output of highly standardized TDM reports. Further studies should examine the reproducibility of LLM-based TDM reports using a larger number of cases, taking a closer look at quality and completeness. In addition, TDM reports should be checked for known risks associated with LLMs, such as hallucinations.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: Use of deepl.com for parts of translation.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Korom R, Kiptinness S, Adan N, Said K, Ithuli C, Rotich O, et al. AI-based Clinical Decision Support for Primary Care: A Real-World Study. https://openai.com/index/ai-clinical-copilot-penda-health/[Accessed 11 Sep 2025].
2. Liu F, Li Z, Zhou H, Yin Q, Yang J, Tang X, et al. Large Language Models Are Poor Clinical Decision-Makers: A Comprehensive Benchmark. In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing [Internet]. Miami, Florida, USA: Association for Computational Linguistics; 2024:13696–710 pp.
3. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025; 642:442–50.
4. Nori H, Daswani M, Kelly C, Lundberg S, Tulio Ribeiro M, Wilson M, et al. Sequential Diagnosis with Language Models. arXiv 2025:2506.22405 https://doi.org/10.48550/arXiv.2506.22405.
5. Kwan HY, Shell J, Fahy C, Yang S, Xing Y. Integrating Large Language Models into Medication Management in Remote Healthcare: Current Applications, Challenges, and. Future Prospects Systems 2025;13:281.
6. Sridharan K, Sivaramakrishnan G. Unlocking the potential of advanced large language model sin medication review and reconciliation: a proof-of-concept investigation. Explor Res Clin Soc Pharm 2024;15: 100492.
7. Poweleit EA, Vinks AA, Mizuno T. Artificial intelligence and machine learning approaches to facilitate therapeutic drug monitoring and model-informed precision dosing. Ther Drug Monit 2023;45:143–50.
8. OpenAI, GPT4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf.[Accessed: 26 Aug 2025].
9. Google, Gemini 2.0 Flash Model Card, https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf.[Accessed 26th august 2025].
10. Wang Y, Ma X, Zhang G, Ni Y, Chandra A, Guoet S, et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. arXiv 2024:2406.01574v6. https://doi.org/10.48550/arXiv.2406.01574.
11. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv 2024: 2312.10997v5. https://doi.org/10.48550/arXiv.2312.10997.
12. Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation, SIGIR '24. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, United States: SIGIR; 2024:2395–400 pp.
13. Hiemke C, Haen E, Paulzen M, Gracia MS, Stegmann B, Singer M, et al. www.psiac.de [Accessed 7 Mar 2025].
14. Hiemke C, Bergemann N, Clement HW, Conca A, Deckert J, Domschke k, et al. Consesus guideline for therapeutic drug monitoring in neuropsachopharmacology: update 2017. Pharmacopsychiatry 2018; 51:9–62.
15. Mainz A. DEGAM guideline S1 053/037. https://www.degam.de/files/Inhalte/Leitlinien-Inhalte/Dokumente/DEGAM-S1-Handlungsempfehlung/053-037%20Medikamentenmonitoring/oeffentlich/S1-HE_Medikamentenmonitoring_Langfassung_201406_mit%20Hinweis%20auf%20Aktualisierung.pdf [Accessed 1 Mar 2023]. Version 1.1.
16. Patsalos PN, Spencer EP, Berry DJ. Therapeutic drug monitoring of antiepileptic drugs in epilepsy: a 2018 update. Ther Drug Monit 2018; 40:526–48.
17. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus 2023;15: e49373.
18. Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. GSM-symbolic: understanding the limitations of mathematical reasoning in large language models. arXiv 2024:2410.05229. https://doi.org/10.48550/arXiv.2410.05229.
19. Pal A, Gema AP, Fourrier C, Minervini P, Ura A. Hugging Face; 2024. https://huggingface.co/openlifescienceai [Accessed 26 Aug 2025].
20. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. MedGemma Technical Report. arXiv 2025: 2507.05201, https://doi.org/10.48550/arXiv.2507.05201.
21. Chen Z, Hernández Cano A, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv 2023:2311.16079, https://doi.org/10.48550/arXiv.2311.16079.
22. Chen J, Cai Z, Ji K, Wang X, Liu W, Wang R, et al. HuatuoGPT-o1 - towards medical complex reasoning with LLMs. arXiv 2024:2412.18925. https://doi.org/10.48550/arXiv.2412.18925.
23. Kim H, Hwang H, Lee J, Park S, Kim D, Lee T, et al. Small language models learn enhanced reasoning skills from medical textbooks. NPJ Digit Med 2025;8:240.