

Views and Opinions

Large language models



Large language models in clinical research: Risks and opportunities

Author: Benjamin V. Ineichen

Affiliations: Assistant Professor and Head of Medical Data Science at the Department of Clinical Research at the University of Bern (UniBE)

doi: [10.54920/SCTO.2026.RAWATCH.11.42](https://doi.org/10.54920/SCTO.2026.RAWATCH.11.42)

Large language models (**LLMs**) are rapidly entering clinical research, with applications ranging from evidence synthesis to protocol development and data extraction. Their ability to process large volumes of text offers efficiency gains. However, recent evidence shows that most evaluations of LLMs are based on simulated or low-complexity tasks, with limited validation in real clinical settings. Only a small fraction of LLM studies use real patient data, and randomised trials on LLMs remain rare. This creates a gap between the perceived capability of LLMs and their actual clinical value. This gap is key for clinical research, where transparency, traceability, and reliability are essential. LLMs should therefore be viewed as assistive tools having clear potential but requiring careful validation and governance before being more broadly integrated into regulated workflows.

The role of LLMs in clinical research

A large language model (**LLM**) is a type of artificial intelligence (**AI**) trained on large amounts of language data to process and generate human language. An LLM can answer questions, summarise documents, and produce written content based on patterns learned from data. LLMs have entered clinical research at a high speed. Since 2022, thousands of studies have explored their use in medicine, covering tasks from literature review to clinical decision support. This growth reflects the strong interest in and clear potential of LLMs.

But what does the evidence from studies involving LLMs actually show? A closer look reveals a mismatch. While the number of studies involving LLMs is large, most of them do not reflect real clinical practice. Many are based on exam-style questions, simulated cases, or curated datasets. Studies using real patient data are far fewer, and prospective trials, including randomised controlled trials (**RCTs**), are rare. The result is a rapidly expanding field with an uncertain evidence base.¹

Clinical research areas where LLMs can make the most impact

Despite these limitations, several areas stand out where LLMs could support clinical research. The first area is evidence synthesis, including systematic review. Screening, extracting, and summarising scientific literature are time-consuming tasks. LLMs can assist in structuring large bodies of evidence and may enable more frequent updates of systematic reviews. For example, LLMs can help in screening large bodies of abstracts for eligibility in systematic reviews or extracting specific data from papers.

The second area comprises clinical trial protocol development and trial design. LLMs can help draft eligibility criteria, summarise prior trials, and identify common endpoints. This may reduce the administrative burden and support more consistent documentation.

The third area that can benefit considerably from LLMs is unstructured text from clinical documentation. Much clinical data (e.g. electronic health records (**EHRs**)) exist in free-text form. LLMs can help convert this information into structured formats that are easier to analyse.

In addition to these core applications, LLMs may support other clinical research tasks. They can assist with drafting or reviewing protocols, informed consent forms, and other regulatory documents – for example by extracting protocol information, checking consistency across documents, or generating first drafts. However, such uses require human review because LLMs may alter required wording, omit information, or introduce factual errors. Studies evaluating LLM-generated clinical trial documents have shown both potential efficiency gains and substantial limitations in regulatory compliance and factual accuracy.

LLMs may also support participant communication. For example, they have been used to convert trial protocols into patient-facing informed consent materials and to answer trial-specific questions using information retrieved from study documents. Other applications include extracting structured information such as eligibility criteria, interventions, and endpoints from trial protocols or other unstructured clinical text; identifying missing or inconsistent information; and structuring protocol content for downstream trial workflows. These applications remain largely experimental and should generally be used as decision or drafting support rather than as substitutes for expert review.

Some applications go further but remain largely exploratory for now. For example, LLMs are discussed in the context of digital twins or trial simulation, where they may help structure patient data or connect it to existing evidence. However, they do not model disease mechanisms and cannot replace predictive or causal models. Other ideas include automated protocol optimisation or real-time decision support during trials. All these use cases are promising, but they raise an important question: how well do these tools perform in real-world conditions?

LLMs' central limitation: Evidence does not reflect clinical reality

The main issue is not what LLMs can do but how they are evaluated. Most studies test LLMs in simplified settings. For example, models are asked to answer multiple-choice questions or interpret short case descriptions. These tasks are easy to standardise and measure, but they are disconnected from real clinical workflows. This evaluation is often called validation: testing whether a model performs reliably for its intended use, ideally on new data and in realistic settings. Because validation sits between model development and deployment, strong performance on a simplified benchmark does not necessarily mean that the model will work reliably in practice.

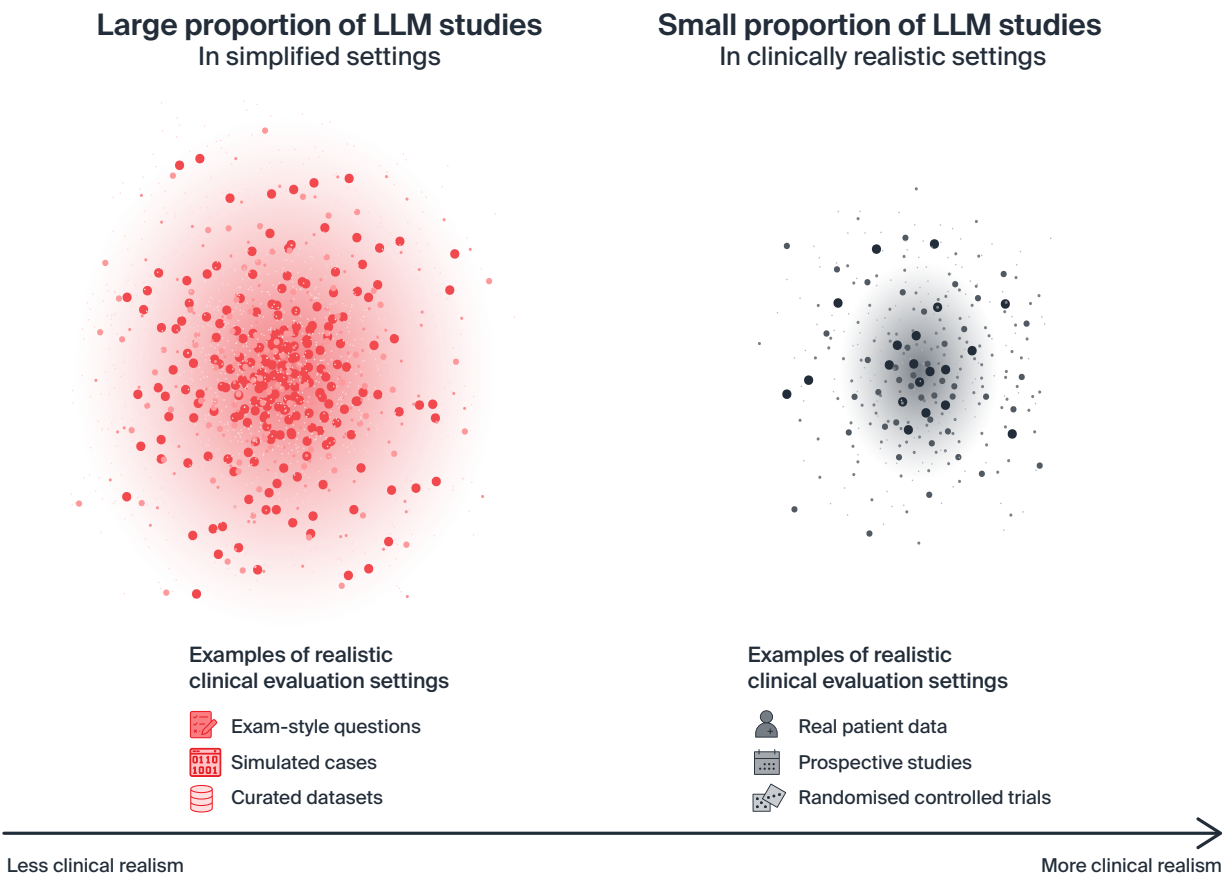
Only a minority of studies use real patient data for the performance evaluation of LLMs.¹ Even fewer studies test LLMs prospectively in clinical environments. LLMs are still rarely evaluated in RCTs, which are often considered the gold standard for evaluating interventions. This matters because performance depends strongly on context. Models often perform well on structured, knowledge-based tasks. But real clinical research involves incomplete data, ambiguity, and complex decision-making. LLMs' performance on simplified tasks may not reflect how they perform when applied to more complex tasks, such as those in real clinical settings. In addition, many studies have a small sample size and/or dataset and lack external validation.¹ This limits confidence in their reported results. Taken together, the current evidence base for LLM performance in clinical research – despite seeming promising at first glance – is not sufficient to support the widespread use of LLMs in clinical settings.

What do performance comparisons really show?

Some studies directly compare the performance of LLMs and clinicians on specific tasks, such as answering medical questions, interpreting clinical cases, making diagnostic decisions, or recommending management. For example, a randomised trial was conducted involving 50 physicians and challenging diagnostic vignettes. GPT-4 alone outperformed the physicians; however, access to GPT-4 did not improve physicians' diagnostic reasoning compared with conventional resources.² At first glance, study results may seem encouraging: in some cases, models perform similarly to or better than humans. But these comparisons need careful interpretation. Performance varies depending on a variety of factors, including the type of task, the extent to which the data reflect real-world clinical practice, and the experience level of the human comparator. LLMs tend to perform best on structured knowledge tasks and less well in realistic clinical scenarios. Moreover, they perform worse when compared to experienced clinicians rather than trainees.

So the key question is not "Can LLMs pass exams?" but rather "Can they support real clinical decisions reliably?" At present, the evidence for such support is limited. **Figure 1** provides a conceptual illustration of this evidence gap based on patterns observed in the literature rather than a systemic synthesis: while many studies assess LLMs in simplified or structured settings, much less evidence comes from real patient data, prospective studies, or randomised controlled trials.

Figure 1: Schematic illustration of large language model (LLM) performance in clinical research



Challenges of using LLMs in a regulated research environment

Clinical research operates under strict regulatory and quality requirements. Data must be accurate, traceable, and reproducible. Decisions must be transparent and auditable. LLMs challenge these principles in several ways. For example, they can generate plausible but incorrect information, and these errors are not always easy to detect. And often, it is not possible to trace outputs to source data, which makes verification difficult. In addition, many widely used models are closed systems, which means that users cannot fully see how a model was trained, what data it was trained on, or how it changes over time. This limits reproducibility, because researchers may not be able to explain why a model produced a specific output or whether the model would produce the same output after an update(s).

There are also concerns about bias, data privacy, and the use of sensitive information. These are not abstract risks in a regulated research environment: they may directly affect data integrity, regulatory compliance, and trust in research outputs.

A pragmatic path forward

So how should clinical research respond to these challenges? LLMs should not be treated as autonomous systems. They are better understood as assistive tools that support specific tasks under human oversight. For example, an LLM may draft eligibility criteria for a clinical trial or extract variables from published studies, but researchers should verify the output against the original protocol or publication before it is used. In addition, a structured approach is

needed: start with well-defined, low-risk use cases; then validate LLM performance on real clinical data and move to prospective studies; and only then consider broader implementation. Furthermore, documentation is essential. Inputs, outputs, and verification steps should be recorded to create an audit trail. Moreover, an LLM should be tested on the actual task and setting in which it will be used. Good performance on general benchmark questions does not necessarily mean it will perform well in a specific clinical research workflow. Most importantly, LLMs should be evaluated in settings that more closely reflect real clinical practice.

Conclusion

LLMs have clear potential to improve efficiency in clinical research. They can support evidence synthesis, documentation, and data structuring. But current evidence on the performance of LLMs is largely based on simplified tasks and in many cases does not reflect real-world use. The key challenge is not technological capability but validation. For now, LLMs should be used cautiously, with clear human oversight and defined roles. Their future in clinical research will depend less on what they can generate – and more on how reliably they can be integrated into rigorous, transparent workflows.

References

- ¹ Chen SF et al. (2026) LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine* 32:1152–1159. doi: [10.1038/s41591-026-04229-5](https://doi.org/10.1038/s41591-026-04229-5)
- ² Goh E et al. (2024) Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open* 7(10):e2440969. doi: [10.1001/jamanetworkopen.2024.40969](https://doi.org/10.1001/jamanetworkopen.2024.40969)

Further reading

- Bean DM et al. (2023) Hospital-wide natural language processing summarising the health data of 1 million patients. *PLOS Digital Health* 2(5):e0000218. doi: [10.1371/journal.pdig.0000218](https://doi.org/10.1371/journal.pdig.0000218)
- Chekroud AM et al. (2024) Illusory generalizability of clinical prediction models. *Science* 383(6679):164–167. doi: [10.1126/science.adg8538](https://doi.org/10.1126/science.adg8538)
- Doneva SE et al. (2024) Large language models to process, analyze, and synthesize biomedical texts: A scoping review. *Discover Artificial Intelligence* 4(1):107. doi: [10.1007/s44163-024-00197-2](https://doi.org/10.1007/s44163-024-00197-2)
- Doneva SE et al. (2025) StudyTypeTeller: Large language models to automatically classify research study types for systematic reviews. *Research Synthesis Methods* 16(6):1005–1024. doi: [10.1017/rsm.2025.10031](https://doi.org/10.1017/rsm.2025.10031)
- Habib AR, Lin AL, and Grant RW (2021) The epic sepsis model falls short: The importance of external validation. *JAMA Internal Medicine* 181(8):1040–1041. doi: [10.1001/jamainternmed.2021.3333](https://doi.org/10.1001/jamainternmed.2021.3333)
- Huang X et al. (2024) A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review* 57, 175. doi: [10.1007/s10462-024-10824-0](https://doi.org/10.1007/s10462-024-10824-0)
- Lieberum JL et al. (2025) Large language models for conducting systematic reviews: On the rise, but not yet ready for use – a scoping review. *Journal of Clinical Epidemiology* 181:111746. doi: [10.1016/j.jclinepi.2025.111746](https://doi.org/10.1016/j.jclinepi.2025.111746)
- Lin A et al. (2025) Large language models in clinical trials: Applications, technical advances, and future directions. *BMC Medicine* 23, 563. doi: [10.1186/s12916-025-04348-9](https://doi.org/10.1186/s12916-025-04348-9)
- Liu F et al. (2025) Application of large language models in medicine. *Nature Reviews Bioengineering* 3(6):445–464. doi: [10.1038/s44222-025-00279-5](https://doi.org/10.1038/s44222-025-00279-5)
- Seyyed-Kalantari L (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* 27:2176–2182. doi: [10.1038/s41591-021-01595-0](https://doi.org/10.1038/s41591-021-01595-0)
- Thirunavukarasu AJ et al. (2023) Large language models in medicine. *Nature Medicine* 29(8):1930–1940. doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)
- Van Calster B et al. (2023) There is no such thing as a validated prediction model. *BMC Medicine* 21, 70. doi: [10.1186/s12916-023-02779-w](https://doi.org/10.1186/s12916-023-02779-w)
- Windecker D et al. (2025) Generalizability of FDA-approved AI-enabled medical devices for clinical use. *JAMA Network Open* 8(4):e258052. doi: [10.1001/jamanetworkopen.2025.8052](https://doi.org/10.1001/jamanetworkopen.2025.8052)