Rev Argent Cir. 2026;118(2):e-ededc- http://dx.doi.org/10.25132/raac.v118.n2.ededc

La importancia de saber preguntar: ingeniería de *prompts* y modelos de lenguaje en la consulta médica

The importance of asking the right questions: prompt engineering and language models in clinical practice

*Enrique Díaz Cantón**

La incorporación de los grandes modelos de lenguaje (LLM) a la práctica clínica ha sido vertiginosa: en pocos meses, pacientes y profesionales comenzaron a consultarlos a diario para informarse, redactar y razonar sobre problemas médicos. El trabajo *Chatbots basados en inteligencia artificial: la importancia de saber preguntar*[1] pone el dedo en una verdad frecuentemente soslayada: la calidad de la respuesta es función de la calidad de la pregunta. Lejos de ser un detalle técnico, esta dependencia tiene consecuencias clínicas directas.

Que el resultado varíe según cómo se formula la consulta no es una intuición, sino un hallazgo reproducible. El estudio de Chen y cols. en *JAMA Oncology* mostró que, ante un mismo escenario clínico planteado con cuatro variantes de redacción, ChatGPT® modificaba sus recomendaciones de tratamiento. El 98% de las respuestas contenía al menos una opción concordante con las guías NCCN, pero el 34,3% incluía, además, alguna recomendación no concordante y el 12,5 % constituía verdaderas "alucinaciones" –terapias inexistentes o curativas para enfermedad incurable–, mezcladas entre consejos por lo demás sensatos y, por ello, difíciles de detectar[2]. La formulación de la pregunta, en suma, se comporta como una variable clínica más.

Ingeniería de *prompts*: una nueva competencia clínica

La ingeniería de *prompts* es el conjunto de técnicas que orientan a un modelo hacia respuestas precisas y verificables. Las más estudiadas incluyen el *zero-shot* (instrucción directa), el *few-shot* (aportar ejemplos resueltos), la asignación de rol, el anclaje con contexto o guías –incluida la generación aumentada por recuperación–y, especialmente, el razonamiento encadenado o *chain-of-thought*, que solicita al modelo desplegar pasos intermedios antes de concluir[3]. Su variante *zero-shot* –el conocido "pensemos paso a paso"– mejora de manera notable el desempeño en tareas de razonamiento complejo[4]. No se trata de retórica: prescribir cómo debe estructurarse la respuesta cambia su exactitud, y por eso revistas de referencia ya publican recomendaciones formales sobre cómo redactar consignas clínicas[5].

¿Qué modelos conviene usar en medicina?

La progresión de los modelos en los bancos de prueba médicos ha sido notable. Med-PaLM 2® alcanzó alrededor del 86,5% en MedQA, un conjunto de preguntas con formato USMLE[6], y Med-Gemini® elevó el estado del arte al 91,1% en 2024[7]. Conviene, sin embargo, una advertencia: el desempeño en exámenes de opción múltiple no equivale a competencia clínica. Las primeras versiones apenas rozaban el umbral de aprobación del USMLE[8], y, como vimos, el mismo ecosistema que sobresale respondiendo preguntas cerradas produce un tercio de recomendaciones no concordantes cuando se lo interroga sobre tratamiento abierto[2]. La tabla 1 sintetiza, con la prudencia debida, esta evidencia heterogénea.

■ TABLA 1

Desempeño de modelos de lenguaje (de generaciones anteriores) en bancos de prueba médicos seleccionados

Modelo (desarrollador)	Banco de prueba	Resultado	Fuente
GPT-3.5 (OpenAI)®	USMLE (Pasos 1-3)	Desempeño cercano al umbral de aprobación	[8]
GPT-3.5 (OpenAI)®	Concordancia con guías NCCN en oncología (mama, próstata, pulmón)	98% con ≥ 1 opción concordante; 34,3% con ≥ 1 no concordante; 12,5 % "alucinaciones"	[2]
Med-PaLM 2 (Google)®	MedQA (estilo USMLE)	Hasta 86,5 % de exactitud	[6]
Med-Gemini (Google)®	MedQA (estilo USMLE)	91,1 % (estado del arte, 2024)	[7]

Nota: los estudios recogidos corresponden a modelos de generaciones anteriores (2023–2024) y emplean metodologías y poblaciones distintas; no constituyen una comparación directa entre modelos ni reflejan el estado del arte actual (véase el apartado siguiente). MedQA: banco de preguntas con formato USMLE; NCCN: National Comprehensive Cancer Network; USMLE: United States Medical Licensing Examination.

*Oncólogo clínico, Instituto Universitario CEMIC, Buenos Aires, Argentina.
Codirector, Programa de Posgrado en Inteligencia Artificial y Medicina, Academia Nacional de Medicina, Buenos Aires, Argentina.
Correspondencia: ediazcanton@iuc.edu.ar

Un panorama en rápida transformación

Conviene subrayar que los estudios reunidos en la tabla 1 emplearon modelos de generaciones anteriores y que el campo avanza a una velocidad inusual. En 2026 conviven modelos de frontera de propósito general notablemente más capaces –como GPT-5.5®, Claude Opus 4.8® o la familia Gemini 3®–y, sobre todo, herramientas especializadas en medicina que anclan cada respuesta en la literatura primaria mediante recuperación aumentada (RAG) y la citan de forma auditable. Plataformas como OpenEvidence® –utilizada por más del 40% de los médicos de los Estados Unidos y con acuerdos de contenido con el *New England Journal of Medicine* y *JAMA*–o Vera Health® –respuestas graduadas según la calidad de la evidencia, en alianza con el American College of Emergency Physicians–ejemplifican esta tendencia[9,10]. La evidencia revisada por pares respalda el enfoque: incorporar conocimiento clínico formal y recuperación semántica elevó el desempeño en el Paso 3 del USMLE a alrededor del 95%, por encima de un LLM nativo (90,5%)[11]. Aun así, ni la mayor potencia ni la especialización eximen de validar cada salida en escenarios clínicos complejos, donde el juicio experto sigue siendo insustituible.

La lección de este trabajo trasciende la anécdota tecnológica. El profesional que incorpore estas herramientas necesitará dos alfabetizaciones complementarias: saber preguntar, esto es, dominar la ingeniería de prompts como una destreza clínica; y saber dudar, sometiendo cada respuesta al juicio experto y a las guías vigentes. Usados con esa disciplina –y con transparencia hacia el paciente–, los *chatbots* pueden ser un aliado valioso. Usados sin ella, multiplicarán los errores con una elocuencia inédita. La pregunta correcta, una vez más, sigue siendo la mitad de la respuesta.

Declaración sobre el uso de inteligencia artificial: *El autor declara que, en la preparación de este editorial, utilizó herramientas de inteligencia artificial generativa basadas en modelos de lenguaje de gran escala como apoyo en la búsqueda y el cotejo de la bibliografía, la organización del texto y la edición de estilo. Todas las afirmaciones, las referencias y los datos cuantitativos fueron verificados por el autor contra las fuentes primarias citadas. El contenido conceptual, las interpretaciones y las conclusiones son de su exclusiva responsabilidad; la inteligencia artificial no figura como autora ni asume responsabilidad alguna sobre el contenido.*

■ ENGLISH VERSION

The integration of large language models (LLMs) into clinical practice has been rapid and significant: within a few months, patients and healthcare professionals began to consult them daily to gather information, draft documents, and discuss medical issues. The paper 'Artificial intelligence-based chatbots: the importance of asking the right questions'[1] highlights a frequently neglected reality: the quality of the answer is contingent upon the quality of the inquiry. Far from being a mere technical detail, this dependency has direct clinical implications.

The fact that the result varies depending on how the query is phrased is not just intuition, but rather a reproducible finding. The study by Chen et al. in JAMA Oncology demonstrated that, when presented with the same clinical scenario described in four different ways, ChatGPT modified the recommended treatment approach. In 98% of responses, at least one NCCN-concordant treatment was included[2]. However, 34.3% also included at least one nonconcordant recommendation, and 12.5% of outputs were hallucinations (non-existent therapies or curative treatments for an incurable disease), which were embedded in otherwise sensible recommendations and thus difficult to detect. In summary, the way in which a question is formulated is an additional clinical variable.

Prompt engineering: a novel clinical competence

Prompt engineering is the set of techniques used to guide a model toward accurate and verifiable responses. The most widely studied techniques include zero-shot (direct instruction), few-shot (providing solved examples), role assignment, context or guideline anchoring—including retrieval-augmented generation—and, notably, chain-of-thought reasoning, which prompts the model to display intermediate steps before arriving at a conclusion[3]. Its zero-shot variant—the well-known "let's think step by step"—significantly improves performance in complex reasoning tasks[4]. This is not a matter of rhetoric: prescribing how a response should be structured alters its accuracy, which is why leading journals are already publishing formal recommendations on drafting clinical prompts[5].

Which models are best suited for use in medicine?

The progress made with models on medical benchmarks has been remarkable. Med-PaLM 2 achieved a score of approximately 86.5% on MedQA, an evaluation dataset of medical knowledge questions from the United States Medical Licensing Examination (USMLE)[6]. In 2024, Med-Gemini achieved a state-of-the-art improvement of 91.1%[7]. Please be aware that

performance in multiple-choice exams is not necessarily an indicator of clinical competence. The early versions barely performed at or near the USMLE passing threshold[8]. As we have seen, the same ecosystem that excels in closed-ended questions produces one-third of non-concordant recommendations when asked about treatments using open-ended questions[1]. Table 1 summarizes this heterogeneous evidence with due caution.

■ TABLE 1

Performance of (previous-generation) language models on selected medical benchmarks.

Model (developer)	Benchmark	Result	Source
GPT-3.5 (OpenAI)	USMLE (Steps 1–3)	Performance at or near the passing threshold	[8]
GPT-3.5 (OpenAI)	Concordance with NCCN guidelines in oncology (breast cancer, prostate cancer, and lung cancer)	98% with ≥ 1 concordant option; 34.3% with ≥ 1 nonconcordant option; 12.5% "hallucinations"	[2]
Med-PaLM 2 (Google)	MedQA (USMLE)	Accuracy up to 86.5%	[6]
Med-Gemini (Google)	MedQA (USMLE)	91.1 % (state-of-the-art, 2024)	[7]

Note: The cited studies pertain to earlier-generation models (2023–2024) and employ different methodologies and populations. They do not constitute a direct comparison between models, nor do they reflect the current state of the art (see the following section). MedQA: benchmark dataset (USMLE subset); NCCN: National Comprehensive Cancer Network; USMLE: United States Medical Licensing Examination.

A panorama under rapid transformation

It is worth noting that the studies compiled in Table 1 utilized previous-generation models, and that the field is advancing at an unprecedented pace. As of 2026, significantly more capable general-purpose frontier models—like GPT-5.5, Claude Opus 4.8, or the Gemini 3 family—coexist with medical-specific tools that anchor every response to primary literature via retrieval-augmented generation (RAG) and cite it in an auditable manner. Examples of this trend include platforms such as OpenEvidence, which is used by over 40% of physicians in the United States and has content partnerships with the New England Journal of Medicine and JAMA, and Vera Health, which provides graded responses based on the quality of evidence in partnership with the American College of Emergency Physicians[9,10]. Peer-reviewed evidence supports this approach: incorporating formal clinical knowledge and semantic retrieval increased Step 3 USMLE performance to approximately 95%, outperforming a native LLM (90.5%)[11]. Nevertheless, the greater power and specialization of these models do not exempt their outputs from validation in complex clinical scenarios, where expert judgment remains irreplaceable.

The lesson of this paper goes beyond the technological anecdote. Professionals integrating these tools will require two complementary literacies: knowing how to ask—that is, mastering prompt engineering as a clinical skill—and knowing how to doubt, subjecting every response to expert judgment and current guidelines. When used with discipline and transparency toward the patient, chatbots can be valuable allies. Otherwise, they will multiply errors with unprecedented eloquence. Once again, the right question is half the answer.

Artificial intelligence disclosure: *The author declares that, during the preparation of this editorial, generative artificial intelligence tools based on large language models were utilized to assist with literature search and cross-referencing, text organization, and copyediting. All statements, references, and quantitative data were verified by the author against the primary sources cited. The conceptual content, interpretations, and conclusions are the sole responsibility of the author; the artificial intelligence is not credited as an author, nor does it assume any responsibility for the content.*

Referencias bibliográficas /References

1. Peña ME, Gigena A, Iglesia F. Chatbots basados en inteligencia artificial: la importancia de saber preguntar. Rev Argent Cir. 2026;118(2):e-1961. DOI: http://dx.doi.org/10.25132/raac.v118.n2.1961
2. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, Bitterman DS. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol. 2023;9(10):1459-62. doi:10.1001/jamaoncol.2023.2954.
3. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-37.
4. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Adv Neural Inf Process Syst. 2022;35:22199-213.
5. Schulte B. Considerations for Prompting Large Language Models. JAMA Oncol. 2024;10(4):538.
6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2.
7. Saab K, Tu T, Weng WH, et al. Capabilities of Gemini models in medicine. arXiv:2404.18416 [preprint]. 2024.
8. 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. doi:10.1371/journal.pdig.0000198.
9. Elkin PL, Mehta G, LeHouillier F, et al. Semantic Clinical Artificial Intelligence vs Native Large Language Model Performance on the USMLE. JAMA Netw Open. 2025;8(4):e256359. doi:10.1001/jamanetworkopen.2025.6359.
10. OpenEvidence. OpenEvidence creates the first AI in history to score 100% on the United States Medical Licensing Examination (USMLE) [comunicado]. 15 ago 2025. Disponible en: openevidence.com.
11. Vera Health (Veracity-Health Inc.). Plataforma de soporte a la decisión clínica basada en evidencia. Disponible en: verahealth.ai (consultado en 2026).