https://doi.org/10.1038/s44298-026-00224-2

Performance of large language models as a source of clinical information on bacteriophage therapy

Nike Walter[1,2] ✉, Derek F. Amanatullah[3], Laurent Debarbieux[4], James B. Doub[5], Tristan Ferry[6,7], Justus Groß[8], Ryszard Międzybrodzki[9,10], Mohammadali Khan Mirzaei[2,7,11,12], Li Deng[2,7,11,12], Kārlis Rācenis[7,13], Gina A. Suh[7,14], Yok-Ai Que[15,16], Andrzej Górski[9,10] & Markus Rupp[2,17]

Bacteriophage therapy is re-emerging as a potential strategy to address antimicrobial resistance, but standardized patient education materials are limited. Large language models (LLMs) are increasingly used for patient-facing medical information. The quality of LLM-generated responses to 20 patient-relevant questions was evaluated by 12 clinicians and research experts in bacteriophage therapy independently rated each response for accuracy, completeness, clarity, and tone/empathy using 5-point Likert scales. Expert suggestions for improvement were recorded. A total of 960 ratings were analyzed. Adjusted mean scores ranged from 3.36 to 3.96 across domains, indicating generally favorable evaluations for all models. Significant differences among LLMs were observed for completeness and tone/empathy (Holm-adjusted $p = 0.042$ for both), but not for accuracy or clarity. Differences were small in magnitude (Cohen's $d = 0.12–0.29$). Claude scored significantly lower than the other models for completeness and tone/empathy, while Perplexity achieved the highest completeness scores. Experts recommended improvements for 34–40% of responses; wrong information was given in 20%. The best responses were revised into an expert-informed patient guide provided as Supplementary Material, presenting a hybrid model in which LLMs generate draft patient information that is subsequently refined by clinical experts, particularly in rapidly evolving therapeutic domains lacking standardized educational resources.

Artificial intelligence (AI) has rapidly spread in the medical field, providing new tools for clinical decision-making and patient communication. Large language models (LLMs) are among the most transformative of these technologies, demonstrating potential across diverse medical domains[1,2]. Initially used for tasks such as data extraction and summarization, LLMs are now increasingly applied to higher-level functions, including diagnostic support and the generation of personalized treatment recommendations[3,4]. In parallel, LLMs have shown promise in direct patient interaction. Their use for drafting responses to patient messages has been associated with greater comprehension and engagement[5] and may contribute to strengthening patient autonomy within the informed consent process[6]. Comparative studies assessing model performance in patient education—such as those focusing on cancer-related inquiries—suggest that LLMs can provide information that is both accessible and clinically relevant[7]. A meta-analysis summarizing these trends concluded that LLMs are already being implemented, at least experimentally, across most medical specialties for a wide range of tasks[8], including diagnostic reasoning[9] and antibiotic stewardship efforts aimed at reducing antimicrobial resistance[10].

Despite this growing body of research, no study to date has examined the role of LLMs in the context of bacteriophage therapy. Bacteriophage therapy, which employs bacteriophages to target specific pathogens, is re-emerging as a potential adjunct or alternative to antibiotics amid the global antimicrobial resistance crisis[11]. The Antibacterial Resistance Leadership Group has emphasized the urgent need for accurate, accessible, and comprehensive communication to guide clinicians considering bacteriophage therapy[12]. The increasing public awareness of bacteriophage therapy has created a growing demand for accessible patient information, while standardized educational materials remain limited. This need is particularly relevant as shared decision-making (SDM) is increasingly adopted in clinical practice, where patients and clinicians jointly consider available treatment options and their potential benefits and harms[13]. In this context, patients increasingly turn to generative AI tools for medical information,

A full list of affiliations appears at the end of the paper. ✉e-mail: nike.walter@ukr.de

Table 1 | Global comparison of LLMs across domains

Domain	*F*	*p* value	Holm-adjusted *p* (primary domains)
Accuracy	1.40	0.241	0.482
Completeness	3.77	0.0105	0.0421
Clarity	0.95	0.414	0.482
Tone/Empathy	3.66	0.0123	0.0421
Overall score (secondary)	2.44	0.0635	—

Fixed-effects regression controlling for rater and question

creating an urgent need to evaluate the quality and safety of such outputs in emerging therapeutic areas. Bacteriophage therapy was deliberately chosen as a case study because it represents a rapidly evolving, high-interest therapeutic area in which standardized patient education resources remain limited and clinical implementation varies across regions. Such characteristics create a setting in which patients may be particularly likely to seek information from generative AI tools. While the present evaluation focuses on bacteriophage therapy, the broader objective is to use this field as a representative model of emerging or translational therapies, where evidence is evolving, regulatory pathways are heterogeneous, and communication needs are substantial. Insights from this study are therefore intended to inform the use of LLMs for patient communication in emerging medical domains more generally.

This study seeks to evaluate LLM-generated responses to clinical and practical questions about bacteriophage therapy and to explore their potential role in supporting evidence-based communication. It aims to examine both the current capabilities and the limitations of generative AI in advancing translational medical communication and to lay the foundation for a practical, trustworthy resource to support the safe and effective communication of bacteriophage therapy in clinical practice. Specifically, the following research questions are addressed: (i) How accurate and complete are LLM responses to clinical questions about bacteriophage therapy? (ii) How clear and comprehensible are these responses for clinical and non-clinical audiences? (iii) How do tone and empathy vary across different models when addressing sensitive therapeutic topics? (iv) What types of improvements do clinical experts identify as necessary for LLM responses to become suitable for real-world application in bacteriophage therapy education and decision support?

Results

Panel characteristics and overall ratings

Twelve experts evaluated responses from four LLMs to 20 phage-therapy frequently asked questions (FAQs), generating 960 individual ratings (12 raters × 20 questions × 4 LLMs). The expert panel had a mean age of 48 years (SD: 7 years). Ten out of twelve raters were male (83%). The expert panel had a mean of 10 years of experience with bacteriophage therapy (SD: 5; range: 4–20 years). No data was missing. Each response was scored on a 5-point Likert scale for accuracy, completeness, clarity, and tone/empathy. All analyses were adjusted for systematic differences in rater severity and question difficulty using regression models including rater and question fixed effects.

Global tests demonstrated significant differences among LLMs in completeness and tone/empathy, but not in accuracy or clarity (Table 1). The overall composite score showed a non-significant trend toward differences between models ($p = 0.064$). Model-adjusted mean scores (scale 1–5) were similar across LLMs (Fig. 1). Across domains, mean scores ranged from 3.36 to 3.96, indicating generally favorable expert evaluations for all models. Perplexity and Gemini tended to achieve the highest adjusted means across domains, whereas Claude consistently showed slightly lower scores. Inter-rater reliability was low across all domains (ICC ≈ 0.02–0.06), indicating substantial variability in scoring thresholds between experts.

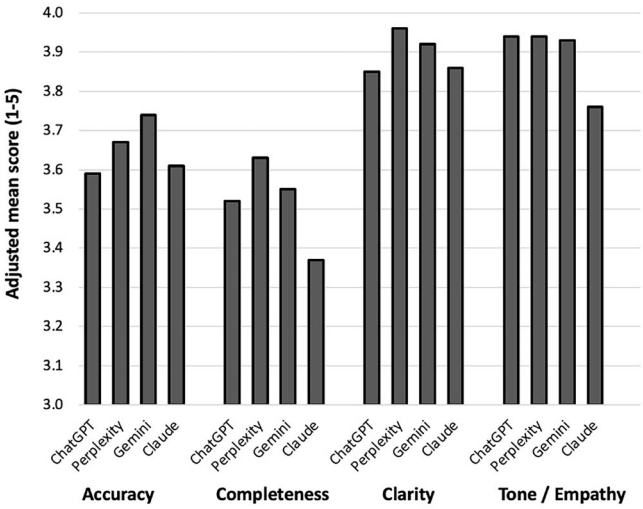

Fig. 1 | Adjusted mean scores (LS means) per LLM across the four evaluated domains (accuracy, completeness, clarity, tone/empathy). Model-based means were averaged over raters and questions on a 1–5 Likert scale. Error bars represent 95% confidence intervals. Between-model differences reached Holm-adjusted statistical significance for completeness and tone/empathy but not for accuracy or clarity.

Between-model differences in completeness and tone/empathy

After adjustment for rater severity and question difficulty, significant differences among LLMs were observed for completeness and tone/empathy (Holm-adjusted $p = 0.042$ for both domains). Perplexity achieved the highest completeness scores, while Claude scored significantly lower than all other models for completeness and tone/empathy (Table 2). No significant differences were observed for accuracy (all Holm-adjusted p values > 0.30) or clarity (all Holm-adjusted p values > 0.50). Observed differences were small in magnitude (Cohen's $d = 0.12$–0.29), corresponding to approximately 15–25% differences in the odds of receiving a higher rating. On a 5-point Likert scale, the largest observed mean difference between models was 0.25 points.

Sensitivity and moderator analyses

Because a 5-point Likert scale is strictly ordinal, a pre-specified sensitivity analysis was conducted using an ordinal (proportional-odds) mixed model with rater and question as random effects; the direction, ranking, and statistical significance of LLM effects were unchanged (Supplementary Table 2). Exploratory analyses including interactions between LLMs and rater characteristics (experience with bacteriophage therapy, years of experience, profession, and sex) showed no evidence of moderation; the relative ranking of LLMs was consistent across rater subgroups. Inter-rater reliability was low (ICC ≈ 0.02–0.06), indicating substantial variability in rater scoring thresholds. This justified the use of regression models controlling for rater effects in all primary analyses.

Improvement suggestions and detection of wrong information

Experts indicated whether each response required improvement. Improvement was suggested for 34–40% of responses. The proportion of responses requiring improvement was lowest for Gemini (34%), followed by Perplexity (38%), ChatGPT (39%), and Claude (40%). Wrong information was detected in 16 responses (20%). When performance was examined at the level of individual questions, the highest-rated response varied across models. Gemini produced the top-rated answer for 8 of 20 questions, followed by Perplexity (5/20), ChatGPT (4/20), and Claude (3/20).

Exploratory analyses including interactions between LLMs and rater characteristics (experience with bacteriophage therapy, years of experience, profession, and sex) showed no evidence of moderation; the relative ranking of LLMs was consistent across rater subgroups. Inter-rater reliability was

Table 2 | Pairwise LLM comparisons for the factor completeness

Comparison	Mean diff	95% CI	Cohen's *d*	Holm *p*
Completeness				
Perplexity vs ChatGPT	+0.10	0.01 to 0.19	**0.12**	0.049
Gemini vs ChatGPT	+0.03	−0.06 to 0.12	0.04	1.00
Claude vs ChatGPT	−0.15	−0.24 to −0.06	**0.18**	0.006
Gemini vs Perplexity	−0.07	−0.16 to 0.02	0.08	0.23
Claude vs Perplexity	−0.25	−0.34 to −0.16	**0.29**	<0.001
Claude vs Gemini	−0.18	−0.27 to −0.09	**0.21**	0.001
Tone / Empathy				
Perplexity vs ChatGPT	0.00	−0.08 to 0.08	0.00	1.00
Gemini vs ChatGPT	−0.01	−0.09 to 0.07	0.01	1.00
Claude vs ChatGPT	−0.18	−0.26 to −0.10	**0.21**	<0.001
Gemini vs Perplexity	−0.01	−0.09 to 0.07	0.01	1.00
Claude vs Perplexity	−0.18	−0.26 to −0.10	**0.21**	<0.001
Claude vs Gemini	−0.17	−0.25 to −0.09	**0.20**	<0.001

Values in bold indicate statistical significance ($p < 0.05$).

low (ICC ≈ 0.02–0.06), indicating substantial variability in rater scoring thresholds. This justified the use of regression models controlling for rater effects in all primary analyses. Experts indicated whether each response required improvement. Improvement was suggested for 34–40% of responses. The proportion of responses requiring improvement was lowest for Gemini (34%), followed by Perplexity (38%), ChatGPT (39%), and Claude (40%). Wrong information was detected in 16 responses (20%). When performance was examined at the level of individual questions, the highest-rated response varied across models. Gemini produced the top-rated answer for 8 of 20 questions, followed by Perplexity (5/20), ChatGPT (4/20), and Claude (3/20).

Development of an expert-informed patient guide

Based on the expert evaluations, the highest-rated response for each question was selected and revised according to the raters' qualitative suggestions for improvement. This iterative process focused on improving clarity, completeness, and tone while preserving scientific accuracy. The resulting set of revised answers represents an expert-informed patient information resource designed to address common questions about bacteriophage therapy (Supplementary Material 1).

Discussion

This blinded expert evaluation demonstrates that LLMs are already capable of generating clinically relevant patient information about bacteriophage therapy, with only modest differences among models. These findings suggest that generative AI may play an emerging role in supporting patient education in rapidly evolving areas of medicine such as bacteriophage therapy. Although our evaluation focused on bacteriophage therapy, the findings may extend to other emerging or rapidly evolving therapies where patients face similar information gaps and may increasingly rely on generative AI tools.

These results align with recent evaluations of LLMs in patient education and clinical communication across diverse specialties. Studies in oncology, cardiology, and chronic disease management have reported that LLM outputs are often rated as accurate and readable, with performance in some settings approaching or occasionally exceeding that of human clinicians for narrow factual questions, while lagging behind in nuanced clinical judgment and contextualization[14–16]. For example, blinded expert assessments have found high accuracy and consistency in ChatGPT responses to disease-specific FAQs, with most ratings in the upper range of Likert scales, although experts still identified gaps in detail and contextual relevance[17]. Conversely, other work comparing ChatGPT to specialist consultants in otorhinolaryngology showed that clinicians outperformed the model in medical adequacy despite the model's high coherence and comprehensibility, highlighting the risk that polished language can mask important clinical omissions or errors[18]. The present study extends this growing body of evidence to bacteriophage therapy, a domain characterized by limited standardized patient materials and complex, evolving evidence regarding indications, formulations, and regulatory pathways.

The observation that differences emerged primarily in completeness and tone/empathy rather than accuracy is notable. This pattern aligns with recent evaluations showing LLMs achieve high factual accuracy (e.g., 89–92% alignment with clinical standards) but exhibit greater variability in completeness and contextual depth, particularly for complex topics. Current LLMs are thus broadly capable of producing factually acceptable responses but vary in how comprehensively and empathetically information is communicated. Patient understanding and trust are strongly influenced by communication quality — a dimension on which LLMs have repeatedly outperformed physicians in blinded empathy ratings — and these dimensions may therefore be particularly relevant for clinical implementation [19–21]. In clinical practice, LLMs could support the development of patient information materials, assist clinicians in preparing educational resources, and help address the growing demand for accessible explanations of emerging therapies. This role may be particularly important given recent evidence that awareness of bacteriophage therapy remains variable even among highly educated populations, while willingness to use the therapy is high. In a recent survey, awareness ranged from approximately 9% to 40%, whereas nearly 85% of respondents expressed readiness to consider bacteriophage therapy. Improving accessible information channels through tools such as LLMs may therefore support informed consent and facilitate SDM for prospective patients[22]. Providing answers to patient questions about bacteriophage therapy can be time-consuming, particularly in small or specialized bacteriophage therapy units where clinicians must balance clinical care, research, and teaching responsibilities. In this context, LLMs could help reduce the informational burden by rapidly generating responses to FAQs, allowing clinicians to focus their time on complex decision-making and individualized patient care. However, expert oversight remains essential, since incomplete or biased outputs may not be apparent to non-expert readers [21,23–25]. Although overall ratings were favorable, experts identified opportunities for improvement in approximately one third of responses. This finding underscores that LLM-generated content should currently be considered a draft or support tool rather than a standalone source of patient education. Importantly, this level of refinement need is consistent with prior work demonstrating that expert review improves safety and clarity of AI-generated medical content, with physicians editing drafts to enhance personalization, empathy, and accuracy[26,27].

Beyond comparative evaluation, the present study enabled the development of a patient-oriented bacteriophage therapy information guide. For each question, the highest-rated LLM response was selected and refined according to expert feedback and suggested improvements. This process resulted in a curated set of answers intended to reflect expert consensus and patient information needs. The guide is provided as Supplementary Material to support clinicians and patients seeking accessible information on bacteriophage therapy. We recognize that the intended audience of this guide is heterogeneous—ranging from health-literate patients and family members through primary-care clinicians unfamiliar with bacteriophage therapy to allied-health colleagues supporting SDM—and that some sections of the current version remain more detailed and technical than may be optimal for patients who are unwell, distressed, or unfamiliar with microbiology. Short lay-language summaries have been added at the start of individual items where feasible, but we consider that a structured consumer review process—including assessment of understandability with validated instruments such as the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), identification of sections requiring simplification, the addition of shorter summaries, visual aids, and clearer signposting—will be required before the guide can be recommended for stand-alone distribution to

patients. However, it has to be considered that only a single response per question was evaluated for each LLM, reflecting typical real-world use but not capturing variability across repeated generations. Also, the expert panel was modest in size ($n = 12$) and predominantly male (83%), and although raters were drawn from 11 centers across Europe and the United States, low- and middle-income settings were not represented. While the panel was composed exclusively of clinicians and researchers with active clinical or scientific engagement in bacteriophage therapy, it did not include perspectives from patients or people with lived experience, from nursing or allied-health colleagues, from health-communication specialists, or from regulatory experts. Each of these perspectives is directly relevant to whether AI-generated information is understandable, trustworthy, actionable, and appropriately caveated for its intended audience, and their absence limits the generalizability of our assessment. Future evaluation should draw on a substantially broader and more diverse review panel. The very low inter-rater reliability observed in this study underscores the subjective nature of evaluating AI-generated patient information; use of rater fixed effects mitigated this variability. In addition, the online sources used to derive the question set are predominantly English-language and skew toward North American and European patient populations, which may under-represent the concerns of patients in other regions and socioeconomic contexts. Also, the study assessed expert perceptions of response quality and did not directly evaluate patient understanding, trust, or clinical outcomes. A follow-up study incorporating a structured consumer reference group is planned. In that study, patients, family members, and representatives from nursing, allied health, health communication, and regulatory affairs will independently evaluate raw LLM outputs and clinician-refined answers along the dimensions of understandability (using PEMAT-P), trustworthiness, tone, perceived usefulness, and preference between the two versions. Such a study would move the field from top-down expert evaluation toward genuinely co-designed, AI-supported patient information—in bacteriophage therapy and, by extension, in other emerging therapeutic areas.

This evaluation was conducted on 5 September 2025 using the free, consumer-facing web interfaces of four generalist LLMs. Since then, the field has moved substantially in two directions of direct relevance to patient information on bacteriophage therapy. First, general-purpose model capability has continued to progress, with successive releases such as Claude Sonnet 4.5, Claude Opus 4.5, and Claude Fable 5, and comparable iterations of the other systems evaluated here. Second, and more consequentially, domain-optimized extensions have emerged. Anthropic has announced the Claude for Life Sciences initiative and Claude Science, a research-oriented AI workbench that integrates specialized biological foundation models via the NVIDIA BioNeMo Agent Toolkit; and Basecamp Research's EDEN family of models, trained on the proprietary BaseData corpus and validated in in vivo antimicrobial peptide-design experiments against WHO priority pathogens, is now accessible through this workbench. These systems are targeted at professional users—researchers, regulatory affairs staff, clinicians and pharmaceutical R&D teams—and are not intended as direct patient-facing tools. They nevertheless mark the emergence of a two-tier LLM ecosystem in which curated, domain-specialized systems may increasingly inform and validate the content that generalist consumer-facing chatbots produce. Our finding that generalist LLMs already generate largely accurate but occasionally incomplete phage-therapy information suggests that the residual completeness and empathy gaps we observed may be addressable, in part, by bridging these two tiers—for example through domain-specific retrieval augmentation, expert-curated reference corpora, and therapy-specific disclaimers embedded at the model provider level.

Our question set was deliberately constructed to bridge two distinct registers of patient enquiry. Community FAQs—those arising in patient forums, social media, lay support groups, and unsolicited enquiries to specialist centers—tend to foreground experiential concerns: access, cost, expected timelines, safety in daily life, and expectation management relative to antibiotics. Clinical FAQs—those raised in physician-patient consultations and SDM—more frequently concern specific indications, contraindications, drug–phage interactions, regulatory pathways, monitoring during therapy, and treatment failure planning. Current generalist LLMs do not explicitly recognize the register of a user's query, and may default to answers pitched at a fixed level of clinical formality regardless of user intent. Evaluating whether next-generation systems can (i) detect the register of an enquiry and (ii) tailor the response—including embedded disclaimers, references to specialist centers, and prompts to consult a treating clinician—is a specific and testable research direction. In the interim, we propose that phage-therapy stakeholder communities engage proactively with LLM providers to place therapy-specific disclaimers and curated reference materials at the model-provider level, particularly for enquiries about access, eligibility, and regulatory status—domains in which factual currency is critical and rapid change is expected. However, responses were generated using publicly available web interfaces, and results may differ across model versions or parameter settings. The LLM landscape is evolving very rapidly. Our data reflect the four consumer-facing systems as they existed on 5 September 2025; performance and safety behavior may change with subsequent updates, and the emergence of domain-optimized extensions (e.g., Claude for Life Sciences, Claude Science, and specialized connectors such as Basecamp Research's EDEN) will likely reshape the professional end of this landscape well before the consumer-facing end.

While we did not compute a formal hallucination rate, wrong and misleading information was detected in 20% of all responses, which underscores the importance of evaluating LLMs for integration into clinical workflows for patient education[28]. LLMs demonstrate promising capabilities for supporting patient communication in bacteriophage therapy. While expert oversight remains necessary, these tools may contribute to improving access to accurate and understandable information in rapidly evolving areas of medicine.

Methods

Identification of FAQs

This study was designed as a blinded cross-sectional expert evaluation of responses generated by four LLMs. FAQs were developed through a four-stage, participatory, patient-informed process[29]. In Stage 1 (source screening, May–July 2025), publicly accessible online resources were reviewed, including the patient-facing informational pages of established phage therapy centers (Eliava Institute, Georgia; Hirszfeld Institute Phage Therapy Unit, Poland; IPATH San Diego, USA; Belgian Coordination Group on Phage Therapy) and lay-language patient forums and social-media discussions (Reddit r/PhageTherapy, Phage Directory community posts, patient-advocacy websites). Enquiries were extracted verbatim and clustered thematically. In Stage 2 (clinician review), the draft list was circulated among clinician co-authors, who cross-referenced it against enquiries received in their respective bacteriophage therapy consultation services in the preceding 12 months; questions raised repeatedly in the clinic but under-represented online (e.g., regulatory access, compassionate-use pathways, drug–phage interactions) were added at this stage. In Stage 3 (patient consultation), the refined list was reviewed with patients who had expressed interest in or undergone bacteriophage therapy at participating centers to confirm face validity and comprehensibility. This consultation was informal in nature: 12 patients ($n = 12$) from participating centers provided input on the wording, ordering, and coverage of the FAQ list, but did not score responses on a validated instrument and were not selected using pre-specified demographic quotas. A subset of these patients additionally reviewed a limited number of clinician-refined answers alongside the corresponding raw LLM outputs and expressed a clear preference for the refined version, describing it as clearer, better structured, and more empathetic; this qualitative feedback was not systematically captured and is reported for transparency only. A structured consumer reference-group evaluation is planned as a follow-up study (see Discussion). In Stage 4 (consolidation), overlapping items were merged, and the list was capped at 20 questions to remain within a rating burden considered feasible for a 12-expert panel. The final set covers the domains of mechanism (Q1, Q2), indications (Q3), safety (Q4, Q5, Q8), administration and treatment course

Table 3 | List of 20 questions presented to the LLMs

Number	Question
1	What is bacteriophage therapy?
2	How does bacteriophage therapy work?
3	What types of infections can bacteriophage therapy treat?
4	Is bacteriophage therapy safe?
5	What side effects or risks should I be aware of with bacteriophage therapy?
6	How are bacteriophages administered?
7	How long does it take to see results with bacteriophage therapy?
8	Can bacteriophage therapy harm my healthy bacteria (microbiome)?
9	What is the regulatory status of bacteriophage therapy internationally (e.g., FDA, EMA, TGA)?
10	How does bacteriophage therapy compare to antibiotics?
11	Can I continue taking my regular medications and antibiotics during bacteriophage therapy?
12	How long does bacteriophage therapy treatment take?
13	Where can I access bacteriophage therapy?
14	How are the bacteriophages selected for my treatment?
15	What happens if bacteriophage therapy doesn't work?
16	What is the cost of bacteriophage therapy?
17	Are there any ongoing clinical trials for bacteriophage therapy I can participate in?
18	Do I qualify for bacteriophage therapy?
19	Can bacteria become resistant to bacteriophage therapy?
20	How are phages stored and prepared?

(Q6, Q7, Q12), regulatory access (Q9, Q13, Q17, Q18), comparison with antibiotics and drug interactions (Q10, Q11), phage selection and personalization (Q14, Q20), treatment failure (Q15), cost (Q16), and resistance (Q19) (Table 3). We acknowledge that this list is not exhaustive. Two clinically important topics—use of bacteriophage therapy during pregnancy and other special populations, and person-to-person transmissibility of therapeutic phages—were identified during post-hoc review as underrepresented in the rating exercise. Rather than modify the pre-specified analysis, we incorporated these topics into the expert-informed patient guide (Supplementary Material 1). Furthermore, the online sources screened are predominantly English-language and skew toward North American and European patient populations; the resulting question set may therefore under-represent the concerns of patients in low- and middle-income settings, and this is addressed as a limitation.

All questions were submitted to publicly accessible AI chatbots on a single date (5 September 2025) to minimize confounding from between-model update cycles. The specific consumer-facing versions queried were: (1) ChatGPT (OpenAI, GPT-5 default web interface); (2) Perplexity AI (Perplexity Inc., default model with real-time web-search enabled); (3) Google Gemini (Google DeepMind, Gemini 2.5 default web interface); and (4) Claude (Anthropic, Claude Sonnet 4 default web interface). All queries were submitted from free-tier, unauthenticated or minimally authenticated accounts, in order to reflect the experience of a lay user seeking information without institutional or premium access. No system prompts, plug-ins, custom instructions, memory features, or connectors were enabled, and default temperature settings were used. We note that the LLM landscape has evolved rapidly since data collection, including the release of Anthropic's Claude for Life Sciences initiative and Claude Science research workbench, of newer general-purpose Anthropic models (Claude Sonnet 4.5, Claude Opus 4.5 and Claude Fable 5), and of specialized biological connectors such as Basecamp Research's EDEN family of models; the implications of these developments for patient-facing information are discussed below. For all systems, the following prompt was used: "*Act as an expert medical doctor who is up to date with the latest scientific research and has years of experience counseling patients with empathy and clarity. Provide comprehensive and easily understandable answers to the following question about bacteriophage therapy! Limit your answer to 200 words.*" Each question was submitted once per LLM in a new chat session. No regeneration or manual editing of responses was performed.

The four LLMs were selected because they represent the most widely accessible and commonly used publicly available generative AI tools at the time of the study. These systems differ in important ways that are relevant to real-world patient use. ChatGPT and Claude primarily generate responses from pretrained model knowledge, whereas Perplexity integrates real-time web search and citation features, and Gemini is tightly integrated into the Google ecosystem and widely accessible via web and mobile interfaces. These differences reflect distinct paradigms of patient interaction with AI tools, including conversational chatbots, search-augmented systems, and ecosystem-embedded assistants. Although usage patterns vary across regions and over time, these tools collectively capture the dominant categories of publicly accessible LLMs that patients are most likely to encounter when seeking health information online. Importantly, these tools are freely accessible and require no specialized technical expertise, making them realistic proxies for patient-initiated medical information seeking

Raters and rating of LLM responses

Each answer was subjected to an evaluation by an independent panel of 12 experts. The panel brought together knowledge and experience in the clinical administration of bacteriophage therapy from 11 centers across Europe and the USA. This international composition was intended to capture diverse healthcare and regulatory contexts. Raters were blinded to the type of LLM; responses were anonymized and presented without identifying the generating LLM. The order of responses was randomized for each rater. Participants were asked to rate each answer with a 5-point Likert scale ranging from 1 ("I strongly disagree") to 5 ("I strongly agree") with the following evaluation criteria:

1. Accuracy: The information is scientifically correct, up-to-date, and reflects current expert consensus or published research.
2. Completeness: The answer addresses all key aspects of the question and provides enough context for understanding.
3. Clarity: The response is clearly written, easy to understand, and uses appropriate language for the target audience.
4. Tone and Empathy: *The tone is respectful, professional, and appropriately empathetic, especially for sensitive or health-related topics.*

Detailed rating anchors for each domain and each Likert score were pre-specified and are provided as Supplementary Table 1. In brief, a score of 5 required all of the following: (i) full agreement with current scientific consensus and no factual errors verifiable against the rater's expert knowledge and, where applicable, guidance from the ESCMID Study Group for Non-Traditional Antibacterial Therapy (ESGNTA); (ii) coverage of all clinically relevant sub-aspects of the question; (iii) language appropriate for a health-literate lay audience without unexplained jargon; and (iv) a professional and empathetic register. A score of 1 indicated a factually incorrect, misleading, or potentially unsafe statement. Because bacteriophage therapy is not the subject of a dedicated Infectious Diseases Society of America guideline, and only limited ESCMID and antimicrobial-stewardship guidance is available, expert consensus—informed by ESGNTA positions in which two panel members hold Education and Clinical Officer roles—served as the primary reference standard. Any factual claim flagged as incorrect by two or more raters was additionally reviewed by the senior authors and cross-checked against the peer-reviewed literature and regulatory documentation (FDA emergency IND and expanded access; EMA reflection papers on phage medicinal products; TGA Special Access Scheme). Such flagged claims are hereafter referred to as "potential hallucinations" and are reported descriptively. No automated hallucination-

detection tool was applied, as none is currently validated for bacteriophage therapy.

The study involved expert evaluation of publicly generated text and did not involve human subject research. Formal ethics approval was therefore not required according to institutional policies.

Statistical analysis

All analyses were conducted at the level of individual ratings. Each answer was rated by multiple experts, and each rater evaluated multiple questions, resulting in a cross-classified data structure with clustering by both rater and question. Ratings in each domain (Accuracy, Completeness, Clarity, Tone/Empathy; 5-point Likert scale) were analyzed using fixed-effects regression models including LLM as the primary predictor and rater and question fixed effects to account for systematic differences in rater scoring severity and question difficulty. The primary model can be expressed as $Y_{irq} = \beta_0 + \beta_1 LLM_i + \alpha_r + \gamma_q + \varepsilon_{irq}$, where Y_{irq} denotes the rating for LLM i, rater r, and question q, and α_r and γ_q represent rater and question fixed effects. Specifically, LLM_i is a categorical variable with four levels (ChatGPT, Perplexity, Gemini, Claude), while rater (12 levels) and question (20 levels) enter as fixed effects to absorb systematic differences in scoring severity and question difficulty; ε_irq is the residual error term. Because a 5-point Likert scale is strictly ordinal, we conducted a pre-specified sensitivity analysis using an ordinal (proportional-odds) mixed model with rater and question as random effects; the direction, ranking, and statistical significance of LLM effects were unchanged (Supplementary Table 2). This approach allows unbiased comparison of LLM performance while using all available ratings without aggregation. For each domain, a global test of the LLM effect was performed using partial F-tests. To control the family-wise error rate across the four pre-specified primary domains (accuracy, completeness, clarity, tone/empathy), Holm adjustment was applied to the resulting p values. An overall composite score (mean of the four domains) was analyzed as a secondary outcome. Adjusted mean scores for each LLM were estimated as least-squares means (LS-means) averaged over the observed distribution of raters and questions. When global tests were significant, pairwise comparisons between LLMs were performed using linear contrasts with Holm correction within each domain. Continuous variables are presented as means with 95% confidence intervals unless otherwise stated. Effect sizes were reported as mean differences on the Likert scale with 95% confidence intervals and standardized effect sizes (Cohen's d) calculated using the model residual standard deviation. Inter-rater reliability was assessed using the intraclass correlation coefficient across LLM–question combinations for each domain. Inter-rater reliability was expected to be modest given the subjective nature of expert evaluations. This was addressed by inclusion of rater fixed effects in all primary models. Exploratory analyses evaluated whether rater characteristics (years of experience, prior administration of bacteriophage therapy, profession, sex) moderated LLM performance by including interaction terms between LLM and rater characteristics. These analyses were considered hypothesis-generating. All statistical tests were two-sided, and a p value < 0.05 after multiplicity correction was considered statistically significant. All analyses were performed using Python (version 3.12; Python Software Foundation, Wilmington, DE, USA) using the packages *statsmodels* and *SciPy*. The dataset and analysis code are available upon reasonable request.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to legal and ethical restrictions on sharing potentially re-identifiable expert rating data, but are available from the corresponding author on reasonable request.

Code availability

Not applicable.

Received: 10 March 2026; Accepted: 11 August 2026;
Published online: 31 August 2026

References

1. Tayyar Madabushi, H. & Jones, M. D. Large language models in healthcare information research: making progress in an emerging field. *BMJ Qual. Saf.* **34**, 73–76 (2025).
2. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med.* **5**, 26 (2025).
3. Alowais, S. A. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
4. Liang, E. N., Pei, S., Staibano, P. & Van Der Woerd, B. Clinical applications of large language models in medicine and surgery: a scoping review. *J. Int. Med. Res.* **53**, 03000605251347556 (2025).
5. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **6**, e379–e381 (2024).
6. Allen, J. W., Levy, N. & Wilkinson, D. Empowering patient autonomy: the role of large language models (LLMs) in scaffolding informed consent in medical practice. *Bioethics* https://doi.org/10.1111/bioe.70030 (2025).
7. Kamminga, N. C. W. et al. Do large language model chatbots perform better than established patient information resources in answering patient questions? A comparative study on melanoma. *Br. J. Dermatol.* **192**, 306–315 (2025).
8. Wei, Q. et al. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J. Biomed. Inform.* **151**, 104620 (2024).
9. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
10. Giacobbe, D. R. et al. Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship. *npj Antimicrob. Resist.* **3**, 14 (2025).
11. Walter, N. et al. The potential of bacteriophage therapy as an alternative treatment approach for antibiotic-resistant infections. *Med. Princ. Pract.* **33**, 1–9 (2024).
12. Suh, G. A. et al. Considerations for the use of phage therapy in clinical practice. *Antimicrob. Agents Chemother.* **66**, e02071-21 (2022).
13. Montori, V. M., Ruissen, M. M., Hargraves, I. G., Brito, J. P. & Kunneman, M. Shared decision-making as a method of care. *BMJ Evid. Based Med.* **28**, 213–217 (2023).
14. Mahmoudi Ghehsareh, M. et al. Expert evaluation of ChatGPT accuracy and reliability for basic celiac disease frequently asked questions. *Sci. Rep.* **15**, 29871 (2025).
15. Su, J. et al. Evaluating large language models for accuracy and completeness of vitiligo patient education: a comparative analysis. *Clin. Cosmet. Investig. Dermatol.* **18**, 2757–2767 (2025).
16. Zhang, Z. et al. Benchmarking large language models against clinicians across hospital levels in cardiovascular decision-making: a cross-sectional vignette-based study. *Sci. Rep.* **16**, 2764 (2025).
17. Ye, Z. et al. An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer. *BMC Women's Health* **24**, 482 (2024).
18. Buhr, C. R. et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. *JMIR Med. Educ.* **9**, e49183 (2023).
19. Vedder, K. et al. Comparative analysis of large language model and physician-generated responses in bariatric patient inquiries: assessing the accuracy and patient satisfaction. *Obes. Surg.* **35**, 3801–3809 (2025).
20. Sorin, V. et al. Large language models and empathy: systematic review. *J. Med. Internet Res.* **26**, e52597 (2024).
21. Li, H., Fu, J.-F. & Python, A. Implementing large language models in health care: clinician-focused review with interactive guideline. *J. Med. Internet Res.* **27**, e71916 (2025).
22. Żaczek, M., Zieliński, M. W., Górski, A., Weber-Dąbrowska, B. & Międzybrodzki, R. Perception of phage therapy and research across

selected professional and social groups in Poland. *Front. Public Health* **13**, 1490737 (2025).
23. Aydin, S., Karabacak, M., Vlachos, V. & Margetis, K. Large language models in patient education: a scoping review of applications in medicine. *Front. Med.* **11**, 1477898 (2024).
24. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
25. Howell, M. D. Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual. Saf.* **33**, 748–754 (2024).
26. Sharma, S., Alaa, A. M. & Daneshjou, R. A longitudinal analysis of declining medical safety messaging in generative AI models. *npj Digit. Med.* **8**, 592 (2025).
27. Johnson, D. et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Preprint at https://doi.org/10.21203/rs.3.rs-2566942/v1 (2023).
28. Agrawal, M., Chen, I. Y., Gulamali, F. & Joshi, S. The evaluation illusion of large language models in medicine. *npj Digit. Med.* **8**, 600 (2025).
29. Mathieson, A., Brunton, L. & Wilson, P. M. The use of patient and public involvement and engagement in the design and conduct of implementation research: a scoping review. *Implement. Sci. Commun.* **6**, 42 (2025).

Acknowledgements

No funding was received for this research. The authors acknowledge the PhageKit project (phagekit.org) and thank the patients at the participating centers for their informal input and helpful discussions on the framing of the question set.

Author contributions

Conceptualization: N.W., M.R. Methodology: N.W., D.F.A., L.Deb., J.B.D., T.F., J.G., R.M., M.K.M., L. Deng, K.R., G.A.S., Y.A.Q., A.G., M.R. Data curation: N.W., D.F.A., L.Deb., J.B.D., T.F., J.G., R.M., M.K.M., L. Deng, K.R., G.A.S., Y.A.Q., A.G., M.R. Investigation: N.W., M.R. Formal analysis: N.W. Writing—original draft: N.W. Writing—review and editing: D.F.A., L.Deb., J.B.D., T.F., J.G., R.M., M.K.M., L. Deng, K.R., G.A.S., Y.A.Q., A.G., M.R. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44298-026-00224-2.

Correspondence and requests for materials should be addressed to Nike Walter.

[1]Department for Trauma Surgery, University Hospital Regensburg, Regensburg, Germany. [2]Institut of Virology, Helmholtz Center Munich - German Research Center for Environmental Health, Neuherberg, Germany. [3]Department of Orthopaedic Surgery, Stanford Medicine, Redwood City, CA, USA. [4]Institut Pasteur, CNRS UMR6047, Bacteriophage Bacterium Host, Université Paris Cité, Paris, France. [5]Division of Clinical Care and Research, Institute of Human Virology, University of Maryland School of Medicine, Baltimore, MD, USA. [6]Department of Infectious Diseases, Hospices Civils de Lyon, Université Claude Bernard Lyon 1, Lyon, France. [7]ESCMID Study Group for Non-Traditional Antibacterial Therapy (ESGNTA), Basel, Switzerland. [8]Department of General and Vascular Surgery, University Medical Center, Rostock, Germany. [9]Bacteriophage Laboratory, Ludwik Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wrocław, Poland. [10]Phage Therapy Unit, Ludwik Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wrocław, Poland. [11]Technical University of Munich, Munich, Germany. [12]Chair of Prevention of Microbial Diseases, Central Institute of Infection Prevention (ZIP), TUM School of Life Sciences, Freising, Germany. [13]Department of Biology and Microbiology, Riga Stradins University, Riga, Latvia. [14]Division of Public Health, Infectious Diseases and Occupational Medicine, Mayo Clinic, Rochester, MN, USA. [15]Department of Intensive Care Medicine, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. [16]Department for BioMedical Research, University of Bern, Bern, Switzerland. [17]Department of Trauma, Hand and Reconstructive Surgery, University Hospital Gießen, Giessen, Germany. ✉e-mail: nike.walter@ukr.de