

Article

Evaluating the Performance of Large Language Models in Evidence-Scarce Scenario: The Diabetic Foot Ulcer Transition Phase [†]

Kamran Shakir ^{1,2} , Hadi Sarlak ^{1,2,*}, Giulia Rogati ² , Alberto Leardini ² , Lisa Berti ^{1,3} and Paolo Caravaggi ² 

¹ Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum-Università di Bologna, Via Zamboni 33, 40136 Bologna, Italy; kamran.shakir2@unibo.it (K.S.); lisa.berti@unibo.it (L.B.)

² Movement Analysis Laboratory and Functional Evaluation of Prostheses, IRCCS Istituto Ortopedico Rizzoli, Via di Barbiano 1/10, 40136 Bologna, Italy; giulia.rogati@ior.it (G.R.); leardini@ior.it (A.L.); paolo.caravaggi@ior.it (P.C.)

³ Physical Medicine and Rehabilitation Unit, IRCCS Istituto Ortopedico Rizzoli, 40136 Bologna, Italy

* Correspondence: hadi.sarlak2@unibo.it; Tel.: +39-051-6366571

[†] A part of this study was submitted to the International Conference on Advanced Technologies & Treatments for Diabetes (ATTD) and was presented as a poster in March 2026.

Abstract

Diabetic foot ulcers (DFUs) impose a substantial burden on people with diabetes and healthcare systems. The post-healing “transition phase” remains clinically challenging with limited guideline support. While large language models (LLMs) are increasingly proposed as clinical decision-support tools, their reliability in evidence-scarce scenarios is largely untested. This exploratory study benchmarked leading LLMs against European clinician consensus for the evidence-scarce scenario of DFU transition-phase management. Six LLMs (ChatGPT-4o, ChatGPT-5.0, Gemini 2.5 Flash, Gemini 2.5 Pro, Claude Sonnet 4.0, and Perplexity) were evaluated for accuracy and hallucination using a two-stage framework. Benchmarks were derived from an online survey of European DFU experts reflecting European-level and national practices (Denmark, Netherlands, UK). Binary outcomes were summarized as proportions with Wilson 95% confidence intervals. Paired within-item comparisons across models were assessed using Cochran’s Q, followed by exact McNemar tests with Holm correction ($\alpha = 0.05$); inferential results were considered supportive due to the limited number of paired items ($n = 8$). Across 192 accuracy assessments and 384 reference checks, LLM accuracy ranged from 50–75% and declined when simulating national practices. Hallucination rates exceeded 50% in several models. LLMs rely on generic recommendations, which contrast with clinicians’ contextual, patient-centered reasoning, suggesting current limitations in their suitability for clinical decision support in DFU transition phase clinical care.

Keywords: generative artificial intelligence; healthcare; large language model; diabetes mellitus; T2DM; diabetic foot ulcers; clinical decision support system



Academic Editor: Jiang Bian

Received: 16 April 2026

Revised: 8 July 2026

Accepted: 16 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The application of artificial intelligence (AI), and in particular of large language models (LLMs) in healthcare, has the potential to be a radical shift leading to a faster and more accurate diagnosis of pathologies, to more focused treatments, and thus to better prognoses [1]. They could also support clinical decision support systems (CDSSs), particularly

in evidence-scarce scenarios in which clinicians must act despite limited high-level evidence. However, this potential of LLMs comes with significant risks, especially in clinical medicine. The most critical is ‘hallucination’, where an LLM produces plausible, confident, and grammatically correct, but factually wrong or fabricated answers [2]. Furthermore, LLMs inherently lack true clinical reasoning and situational awareness, as they do not understand a patient’s unique context, cultural preferences in medical practices, or the constraints of a local healthcare system [3]. This may raise serious concerns about their safety and the reliability of LLM-generated content in evidence-scarce scenarios. Thus, rigorous evaluations of current LLMs are essential before integrating them into real-world clinical scenarios.

One of the fastest-growing systemic pathologies worldwide, diabetes mellitus (DM) represents an ideal application area for AI and LLMs. DM is a modern-day pandemic that significantly contributes to population morbidity, mortality, and healthcare costs [4]. Among the several consequences of DM, people affected by peripheral neuropathy have a higher chance of developing diabetic foot ulcers (DFUs), which are a leading cause of non-traumatic lower extremity amputations [5]. DFU management requires intensive, long-term, multidisciplinary care, which places a significant burden on healthcare systems [6]. Plantar offloading is a major strategy for preventing DFUs and promoting healing after ulceration in at-risk foot regions [7–9]. Research on current industry practices indicates that manufacturers and clinicians alike prioritize offloading effectiveness as a key functional requirement for diabetic footwear [10]. For the optimal plantar offloading, custom-made therapeutic footwear is recommended, as these devices have been shown to reduce the recurrence rate of DFUs [11]. While current clinical practice aims to treat active ulcers until wound closure, this may not be a robust indicator of healing, but rather only one phase in preventing DFU recurrence. In fact, the newly healed scar tissue is structurally and functionally weaker than the corresponding healthy soft tissue [12], making it vulnerable to DFU recurrence. It has been estimated that around 40% of DFU recurrences occur within the first year of ulcer healing, while 65% occur in the following 3–5 years [6]. This brief but critical period, immediately following ulcer closure and lasting a few weeks till the skin and underlying tissue have completely recovered, is called the transition phase [12]. During this vulnerable period, a person with a DFU transitions from a highly protective offloading device (e.g., a total-contact cast or walker boot) to long-term therapeutic footwear for secondary prevention.

Despite its clinical importance, the transition phase remains an under-investigated subject in DFU clinical care [7,12]. While extensive practical guidelines, such as those published by the International Working Group on the Diabetic Foot [13], provide detailed recommendations for managing active ulcers and for long-term prevention in at-risk patients, they offer little specific guidance for the transition phase. In fact, the heterogeneity of patient demographics, ulcer locations, and healing trajectories makes it difficult to design and execute large-scale randomized controlled trials that could inform with high-level evidence [14]. In this scenario, the collective experience and judgment of experienced clinicians become the de facto standard of care (Level V evidence) [15].

Therefore, in clinical scenarios where formal guidelines are lacking, LLMs could serve as clinical decision support systems. Since LLMs are trained on very large datasets, such as medical textbooks, journal articles, and clinical case reports, this study aims to determine whether current LLMs can synthesize patterns and implicit knowledge that align with expert consensus and thus may serve as a proxy for Level V evidence.

The primary objective of this study was thus to systematically evaluate the accuracy and reliability of publicly available and commonly used LLMs in providing clinical guidance for the DFU transition phase. Accuracy was evaluated by how well LLM responses

aligned with the clinicians' consensus, whereas reliability was determined by the frequency of hallucinated references and the accuracy of the responses. Since clinical practice varies significantly across countries due to differences in funding, regulations, available technologies, and cultural practices, the secondary objective was to assess the models' ability to adapt their responses to specific geographic and local healthcare systems.

2. Materials and Methods

To test the accuracy of different LLMs in guessing the clinical consensus on several questions regarding the transition phase of people with DM at risk of DFU, a two-stage methodology was adopted. First, a benchmark was established on the offloading recommendations for the transition phase. This was obtained using a web-based survey administered to senior clinicians working in Europe. Six commonly used LLMs were tested for accuracy and response hallucination against the benchmark on the same offloading questions. A schematic of the study workflow is shown in Figure 1.

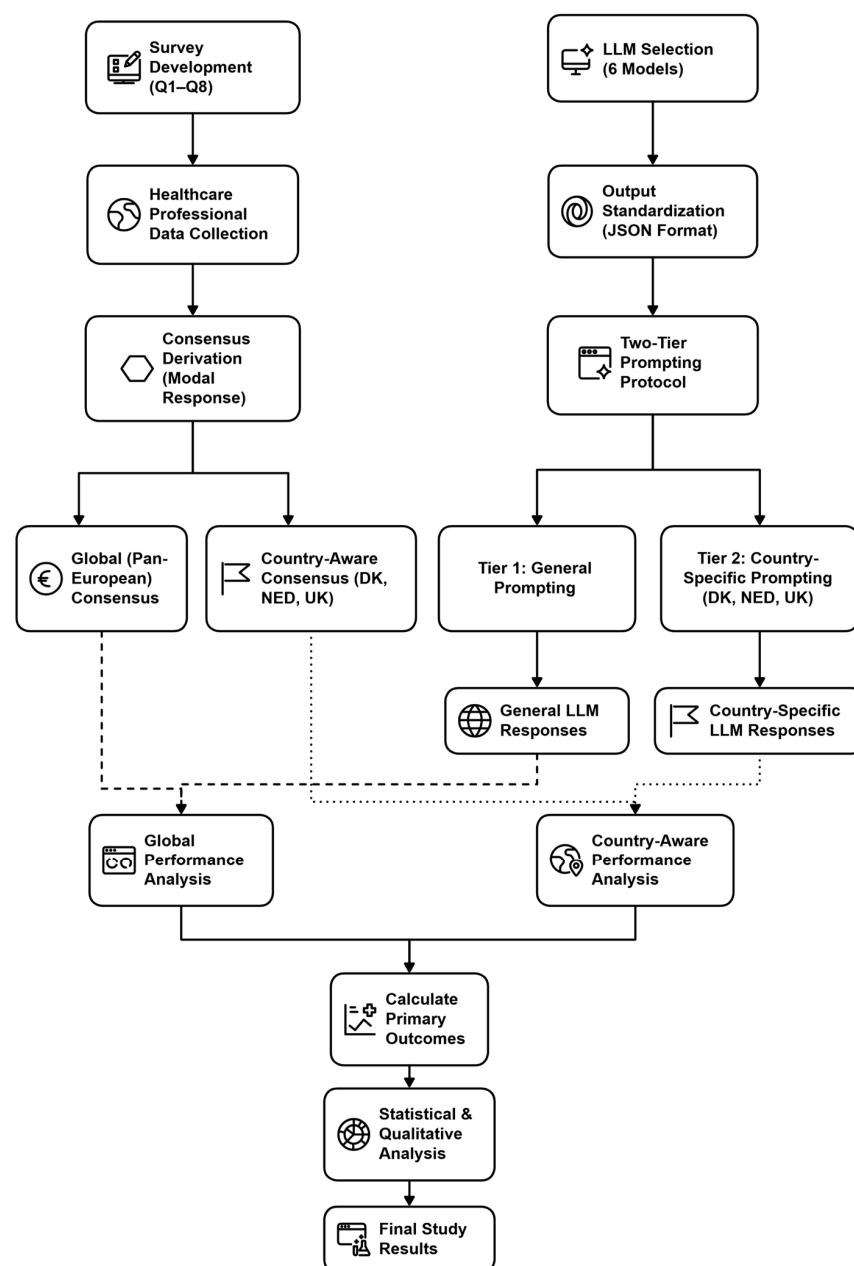


Figure 1. Schematic of the study protocol.

2.1. Benchmark Establishment: The Healthcare Professionals Survey

2.1.1. Survey Characteristics

A questionnaire-based survey was developed focusing on “Footwear solutions for individuals in the transition phase” and implemented via a General Data Protection Regulation (GDPR)-compliant web-based application. It consisted of three demographic questions, seven multiple-choice questions, and one descriptive question (Appendix A). This survey was designed to evaluate the characteristics of the offloading modalities to be implemented during the transition phase and comprised four key topics: the timeframe for offloading; features of the insoles; features of the midsoles; and features of the upper.

2.1.2. Clinician Cohort and Consensus Derivation

The questionnaire was distributed to healthcare professionals specializing in diabetic foot care (podiatrists, physicians, orthotists/prosthetists, and shoe technicians) across multiple European countries through a QR code with a link to a SurveyMonkey questionnaire (<https://www.surveymonkey.com>) during the Diabetic Foot Study Group (Prague, Czech Republic, 2025) and personal referrals from May to September 2025. Two consensus datasets were derived from their responses:

1. **Global (Pan-European) Consensus:** Responses from all survey participants were used. For multiple-choice items (Q1–Q7), the modal response was designated the consensus benchmark. Thematic analysis was performed for Q8, and the modal theme was considered as the benchmark.
2. **Country-Aware Consensus:** For Denmark, the Netherlands, and the United Kingdom, countries with sufficient respondent numbers ($n \geq 4$), modal responses were calculated separately, giving indicative national trends.

A modal-response approach was selected to draw consensus among the healthcare professionals. Consensus strength was calculated as the ratio of the respondents choosing the modal response to the total number of responses for a given question. For Q8, open-text responses were first coded into recurring themes (Availability/Variety; Adaptability/Customization; Accessibility/Cost; Timeliness/Speed; Standardization/Guidelines; Quality/Aesthetics), and then the modal theme was identified.

$$\text{Consensus strength (\%)} = \frac{\text{Number of responses selecting the modal answer (n)}}{\text{Total valid responses for that question (t)}} \times 100 \quad (1)$$

2.2. LLMs Selection and Analysis

Six publicly available and widely used LLMs at the time of the study were selected: ChatGPT 4-o (OpenAI); ChatGPT 5.0 (OpenAI); Gemini 2.5 Flash (Google); Gemini 2.5 Pro (Google); Claude Sonnet 4.0 (Anthropic), and Perplexity Sonar (Perplexity AI). A two-tier zero-shot prompting protocol was designed to assess the models’ performance in both global and country-specific contexts. Models were accessed through their publicly available web-based interfaces between August and September 2025 in a new window with cookies removed. For each model, the standard graphical user interface available at the time of testing was used. No external documents, uploaded files, custom retrieval-augmented generation database, or few-shot examples were provided. The default settings of their respective web-based applications were used for all models. ChatGPT 4-o (OpenAI), Gemini 2.5 Flash (Google), Claude Sonnet 4.0 (Anthropic), and Perplexity (Perplexity AI) were free-to-use, whereas ChatGPT 5.0 (OpenAI) and Gemini 2.5 Pro (Google) were paid versions at the time of the study. The complete prompting protocol and its rationale are reported in Appendix B. To systematically compare the responses, all prompts required models to return answers to (Q1–Q8) in the following predefined JSON format. This

eliminated variability in structure and phrasing, enabling objective benchmarking against clinician consensus data.

```
{
  "answer": "<one of the allowed options>",
  "explanation": "<1–3 lines why this is clinically appropriate>",
  "references": ["<Author, Year, Journal>", "<Author, Year, Journal>"]
}
```

2.3. Study Outcomes and Data Analysis

The study assessed two primary metrics of LLM performance: accuracy and hallucination rate. Accuracy was defined as a binary outcome. A model's response was coded as accurate (score = 1) if it exactly matched the consensus response for Q1–Q7 and unambiguously followed the same theme in Q8 as that of the consensus; all deviations were coded as inaccurate (score = 0). The hallucination was defined as the presence of fabricated or unverifiable scientific references. Real but irrelevant references, references that did not support the specific model-generated statement, and quote-level claim verification were not coded as separate quantitative outcomes in the present analysis. If both the references returned by a model were not hallucinated, its responses were coded as 0, otherwise it was encoded as 1. Reliability was calculated as a combined descriptive index incorporating both answer accuracy and reference validity. The formulas used to calculate accuracy (%), hallucination (%), and reliability (%) are reported in Equations (2)–(4):

$$\text{Accuracy}(\%) = \frac{\sum \text{Correct responses}}{\text{Total responses}} \times 100 \quad (2)$$

$$\text{Hallucination}(\%) = \frac{\sum \text{Responses with } \geq 1 \text{ fabricated/unverifiable reference}}{\text{Total responses}} \times 100 \quad (3)$$

$$\text{Reliability}(\%) = \text{Accuracy}(\%) \times \left[1 - \frac{\text{Hallucination rate}(\%)}{100} \right] \quad (4)$$

Binary outcomes (accuracy and hallucination) were summarized descriptively as proportions with Wilson 95% confidence intervals. For paired, within-item comparisons across the six models addressing the same eight questions, overall differences were assessed via Cochran's Q. When significantly different, pairwise differences between models were tested using the exact McNemar test; Holm's correction was applied for multiple comparisons. Due to the small number of paired items ($n = 8$), inferential results were considered only as supportive of the descriptive analysis ($\alpha = 0.05$). Therefore, p -values were interpreted as supportive rather than definitive evidence of model performance. Statistical analysis was performed using MATLAB (R2025b, MathWorks, Natick, MA, USA) and IBM SPSS Statistics (v27.0, IBM Corp., Armonk, NY, USA). Survey responses to Q8 were analyzed thematically to identify recurrent themes (e.g., short time duration, technology integration, patient education). LLM response themes were compared with those derived from clinician responses to assess accuracy.

3. Results

3.1. Survey

Out of the 16 professionals who completed the survey, 6 were physicians; 7 were podiatrists; 2 were shoe technicians, and 1 was a researcher. Around 57% of the respondents had 11–20 years of experience working with people with DFUs, and around 20% had more than 20 years of experience in the field. In terms of geographic location, six worked in the Netherlands, five in the United Kingdom, four in Denmark, and one in Italy. The consensus

of the professionals is reported in Table 1. Country-wise consensus on the characteristics of transition-phase footwear for people with DFUs is reported in Table 2. Due to the limited sample size per country ($n = 4$ to 6), these results should be interpreted as an exploratory expert-consensus comparator rather than a definitive clinical standard. Given that the LLMs were required to choose one of the same predefined options presented to healthcare professionals, the purpose here was to test alignment with the dominant practice pattern in an evidence-scarce clinical scenario.

Table 1. Healthcare professionals' consensus (global) on the characteristics of transition-phase footwear for people with DFUs. Consensus strength was calculated as the number of respondents selecting the modal response divided by the total number of valid responses recorded for that question, multiplied by 100.

Question	Consensus (Modal Response)	Total Responses Recorded for This Question (t)	Number of Respondents Selecting the Modal Response (n)	Consensus Strength (%)
Q1: Time to Long-Term Footwear	1–3 months	15	11	73.3
Q2: Footwear Used in Transition	Other (offloading devices/boots)	15	8	53.3
Q3: Special Attention Needed	Yes	16	14	87.5
Q4: Recommended Shoe Upper Type	Closed (like sneakers)	15	12	80.0
Q5: Recommended Insole Type	Custom-made	11	9	81.8
Q6: Recommended Midsole/Outsole Material	Semi-rigid	13	8	61.5
Q7: Recommended Midsole/Outsole Shape	Forefoot rocker	13	8	61.5
Q8: Most Requested Improvement	Availability/variety	11	4	36.4

Table 2. Healthcare professionals' consensus (country-wise) on the characteristics of a transition-phase footwear for people with DFUs.

Question	Consensus (Denmark)	Consensus (Netherlands)	Consensus (UK)
Q1: Time to Long-Term Footwear	3–6 months	1–3 months ^β	1–3 months ^β
Q2: Footwear Used in Transition	Wear off-the-shelf therapeutic footwear	Other (made temporary shoe) ^β	Other (a removable cast) ^β
Q3: Special Attention Needed	Yes ^β	Yes ^β	Yes ^β
Q4: Recommended Shoe Upper Type	Closed (like sneakers) ^β	Closed (like sneakers) ^β	Closed (like sneakers) ^β
Q5: Recommended Insole Type	Custom-made ^β	Custom-made ^β	Custom-made ^β
Q6: Recommended Midsole/Outsole Material	Semi-rigid ^β	Rigid	Semi-rigid ^β

Table 2. Cont.

Question	Consensus (Denmark)	Consensus (Netherlands)	Consensus (UK)
Q7: Recommended Midsole/Outsole Shape	Forefoot rocker ^β	Forefoot rocker ^β	Roller design
Q8: Most Requested Improvement	More accessible and variety	Availability/Variety ^β	Quality/Aesthetics

^β represents those values that also match the global consensus.

3.2. Overall Performance of LLMs Against Global (Pan-European) Clinician Consensus

A total of 48 responses (6 LLM models × 8 questions) were compared with the global consensus. Ninety-six references reported by the LLMs (6 LLM models × 8 questions × 2 references for each question) to support their responses were analyzed to determine the hallucination rate. Model-wise accuracy and hallucination rates are reported in Table 3. Performance varied considerably across models, with a mean accuracy of 62.5% (range, 50.0–75.0%). No statistical difference in accuracy was found (Cochran’s Q, $p = 0.587$). In contrast, models differed significantly in hallucination (Cochran’s Q, $p = 0.0127$), with a mean hallucination rate of 56.25% (range, 12.5–100%). Post hoc pairwise comparisons between models using the McNemar test with Holm correction across the 15 pairs resulted in no significant differences after adjustment ($\alpha = 0.05/15 \approx 0.0033$). The two paid LLMs employed in the study, Gemini 2.5 Pro and ChatGPT 5.0, achieved the best accuracy rates of 75%. However, this should not be interpreted as evidence that paid access can be associated with superior performance. The study was not designed to isolate the effect of subscription tier from other factors such as model architecture, training updates, retrieval behavior, or interface settings. Claude Sonnet 4.0 balanced accuracy with a lower hallucination rate and was also free to use. None of the models used in the study achieved 100% accuracy. Responses from all LLMs matched the consensus for Q3, Q4, and Q7. Question-wise LLM performance is reported in Appendix C.

Table 3. LLMs’ accuracy and hallucination rate when compared with the global consensus. For each model, accuracy was calculated as the proportion of questions for which the model response matched the clinician-consensus benchmark. Hallucination rate was calculated as the proportion of model responses in which at least one returned reference was fabricated or unverifiable. Reliability was a combined metric of accuracy and hallucination, defined as the product of accuracy and (1 minus the hallucination (%)) divided by 100), expressed as a percentage.

LLM	Accuracy (%) ^δ 95% CI-Wilson	Hallucination (%) ^θ 95% CI-Wilson	Reliability (%) ^δ
Claude Sonnet 4.0 (free)	62.5 (30.6–86.3)	12.5 (2.2–47.1)	54.7
Gemini 2.5 Pro (paid)	75.0 (40.9–92.9)	50.0 (21.5–78.5)	37.5
ChatGPT 4-o (free)	50.0 (21.5–78.5)	37.5 (13.7–69.4)	31.3
ChatGPT 5.0 (paid)	75.0 (40.9–92.9)	62.5 (30.6–86.3)	28.1
Perplexity (free)	62.5 (30.6–86.3)	75.0 (40.9–92.9)	15.6
Gemini 2.5 Flash (free)	50.0 (21.5–78.5)	100.0 (67.6–100.0)	0

^δ indicates higher is better; ^θ indicates lower is better.

A trend in LLMs' ability to provide the right answer but fail to provide a correct scientific reference to support it was observed. Gemini 2.5 Flash provided the weakest results, showing the lowest accuracy (50%) and failed to provide any valid reference to its responses across all questions.

3.3. The Influence of Geographic Context on AI Model Performance

A total of 144 answers (6 LLM models $\times$ 8 questions $\times$ 3 countries) given by the LLMs were compared with the consensus of the respective countries to verify the accuracy. The LLMs also returned 288 references (6 LLM models $\times$ 8 questions $\times$ 3 countries $\times$ 2 references), which were analyzed for hallucination rates. Model-wise accuracy and hallucination rates are reported in Table 4. Within each country, Denmark, the Netherlands, and the United Kingdom, Cochran's Q indicated no accuracy differences among models (pDK = 0.208; pNED = 0.292; pUK = 0.980) but significant differences in hallucination (pDK = 0.004; pNED = 0.002; pUK < 0.001). Post hoc pairwise comparisons using the exact McNemar test with Holm's correction found no significant pairwise difference for either outcome in any country. When instructed to simulate a clinician's practice in the above-mentioned countries, some LLMs struggled to adapt to specific national healthcare contexts, and a trend toward higher mean hallucination rates was observed. Perplexity achieved the highest mean accuracy, but at the cost of a high hallucination rate. ChatGPT 5.0 exhibited a significantly lower hallucination rate than its global performance while maintaining a mid-range accuracy. It appears that most LLMs performed better than their global average (pooled across models and questions) for Denmark, suggesting that the Danish benchmark was the easiest for the LLMs to align with.

Table 4. Accuracy and hallucination rates of the 6 LLM models in country-aware scenario; β indicates an accuracy equal to or higher than the global average for the respective LLM.

LLM	Accuracy (%)				Hallucination (%)			
	DK	UK	NED	Mean	DK	UK	NED	Mean
ChatGPT 5.0 (<i>p</i>)	75.0 β	50.0	50.0	58.3	25.0 Φ	0.0 Φ	0.0 Φ	8.3 Φ
Gemini 2.5 Pro (<i>p</i>)	50.0	50.0	75.0 β	58.3	62.5	87.5	75.0	75.0
Claude Sonnet 4.0 (<i>f</i>)	62.5 β	50.0	50.0	54.2	25.0	37.5	37.5	33.3
Perplexity (<i>f</i>)	87.5 β	50.0	75.0 β	70.8 β	87.5	87.5	87.5	87.5
ChatGPT 4-o (<i>f</i>)	75.0 β	50.0 β	50.0 β	58.3 β	87.5	87.5	87.5	87.5
Gemini 2.5 Flash (<i>f</i>)	62.5 β	37.5	50.0 β	50.0 β	87.5 Φ	37.5 Φ	71.4 Φ	65.5 Φ

Φ indicates the hallucination rate being equal to or lower than the global average for the respective LLM; (*p*): paid version, (*f*): free version.

4. Discussion

The use of AI/LLMs in DFU care has primarily focused on medical knowledge extraction [16,17], prognosis [18,19], and the fact-checking of established clinical recommendations [20]. This study, to the best of the authors' knowledge, is the first of its kind to assess the performance of GenAI models in an evidence-scarce DFU care setting. This study has highlighted that for this specific clinical subject, model accuracy tends to be higher for paid LLMs. GenAI chatbots can generate different responses to the same healthcare-related questions based on the user's location and platform [21], and this study has further highlighted that the geolocation appropriateness of the responses differs largely from model to model. The LLMs tested here are generalists by design and were trained on vast datasets and textual information available online [22]. Although these models performed well in a global context, they appeared less reliable when addressing questions in country-specific contexts, owing to a high rate of hallucination.

Medical practice is deeply rooted in a complex ecosystem of local and international guidelines, insurance reimbursement policies, the availability of specific medical devices and pharmaceutical products, referral pathways, and most importantly, the unwritten experience of the healthcare professional, often passed down through mentorship [23,24]. This can be observed from the clinician consensus. For example, in the Netherlands, healthcare professionals favor a 'rigid' midsole (Table 2) for transition-phase footwear, while the global consensus is to prescribe semi-rigid midsoles. This decision may be driven by the local availability of certain footwear materials or by favorable clinical outcomes with rigid materials. The LLMs, lacking access to this layer of real-world awareness, defaulted to the more globally common recommendation of 'semi-rigid', resulting in a contextually inaccurate answer. Similarly, healthcare professionals in Denmark prefer custom-made insoles and rigid midsoles for the transition phase footwear for people with DM, perhaps due to the government subsidizing bespoke footwear [25]. Models such as Claude Sonnet 4.0 and Gemini 2.5 Flash overlooked this particularity and recommended semi-custom or prefabricated alternatives, which provide less offloading than custom-made counterparts [26] but are more cost-effective, aligning with generalized European recommendations. The failure of LLMs to adapt to country contexts may not be due to a knowledge gap that can be filled with more data; rather, it may reflect a failure of training on relevant, context-aware data, which is not easy to gather in textual form. This is supported by the fact that LLMs, when aided by retrieval-augmented generation using curated local guidelines and institutional protocols, have been shown to produce more clinically accurate responses and fewer hallucinations [27]. A true clinical AI is required to move beyond the limitations of its generalized training set and should adapt to the constraints of the specific environment in which care is provided. The results of this study suggest that the current generation of LLMs may face significant challenges in performing reliably in the clinical scenario of the DFU transition phase.

Apart from the varying accuracy of the models for country-aware scenarios, the LLMs also suffered from high rates of hallucination. In sectors that directly affect human life and well-being, such as the medical industry, the validity of knowledge is of the utmost importance. A clinical recommendation by a healthcare professional is only as good as the evidence on which it is based. The high rates of hallucination recorded in this study represent a critical failure of this principle. What is even more concerning is that the LLMs provided correct answers but cited non-existent or irrelevant references, thus generating a phenomenon of unverifiable accuracy. Other researchers have also assessed this phenomenon and found that 50–90% of LLM responses are not fully supported by the sources [28]. The analysis of the responses of the LLMs for the open-ended Q8 revealed their ability to generate recommendations that are highly plausible using medical terminology. Though these answers (reported in Appendix C) may seem sophisticated and correct, they often lack a genuine clinical insight. This results in a subtler and harder-to-detect risk of adopting suboptimal or inappropriate strategies simply because they are presented in a professional-sounding manner [29]. For example, qualitative analysis of the same open-ended Q8 from the clinicians' survey revealed patient-centric issues, such as aesthetics and costs, and a rejection of standardized protocols for the transition phase, as it is case-dependent. This risk aligns with well-documented evidence of automation bias in CDSS, in which convincing but incorrect responses may mislead clinicians [30,31]. Replacing experience-driven decision-making with generic, algorithmically generated statements by LLMs may risk deteriorating the quality of care.

The generation of LLMs used in the study may not be considered reliable under clinical uncertainty, even though some models achieved higher accuracy than others. The reliability index of the paid models was not always better than that of the free-to-use models. LLMs like Perplexity prioritize plausible-sounding text; others (like Claude Sonnet) adhere better to known information. This is mainly because of their tendency to hallucinate, especially in context-specific country scenarios, which could make them even more dangerous in evidence-scarce domains. Unsupported claims appear in LLM responses even when citations are provided, as found in this study and others [17]. Some studies have also reported on the ethical and accuracy concerns of using general LLMs in clinical settings [22,32]. When clear guidelines exist, an LLM's output can be more easily verified. However, in the absence of clearly defined guidelines, the user can only rely on the models' explanations and references to fact-check the recommendation. If these references are hallucinations, it undermines trust precisely when external verification is most needed. Nevertheless, this should not imply that LLMs lack clinical utility. Instead, their contribution could be of a carefully supervised cognitive tool for specialists. An experienced clinician, fully aware of LLMs' limitations, could use them as a differential diagnostic tool. This way, LLMs could provide possibilities, but the guarantee of validity lies with the human expert and their clinical judgment to either accept or reject them. This aligns with the broader consensus that AI should assist, rather than replace, human intelligence in critical fields [33,34]. The clinician's expertise will discard the plausible but impractical suggestions and ground the AI's abstract output in the concrete reality of the individual patient in evidence-scarce scenarios.

This study benefits from certain strengths. Its novel design, directly benchmarking LLMs against practicing clinician consensus, provides a more valid assessment of clinical utility than standardized tests or fact-recall evaluations. The resulting distribution of the survey respondents reflects the multidisciplinary nature of DFU management, in which physicians, podiatrists, footwear specialists, orthotists/prosthetists, shoe technicians, and researchers may all contribute to offloading decisions and footwear prescription. The focus on an evidence-scarce scenario is a key contribution, as it is in these areas of uncertainty where decision support is most needed and where the risks of AI are most pronounced. The current study sought to evaluate the performance of these AI models by asking them questions that may not have straightforward, literature-supported answers. The use of a two-tiered, country-specific prompting protocol is a major strength, yielding insights into the critical challenge of contextual adaptation for AI in complex clinical settings.

The current study is not without limitations. The field of AI is evolving at an extraordinary pace, and the specific models tested in this cross-sectional study represent only a snapshot in time; future iterations may exhibit improved performance. The clinician panel, while representing a consensus across multiple European countries, was small and may not capture the full spectrum of clinical practice in Europe or worldwide. The consensus was derived from the modal response, which may omit significant minority opinions. Since a formal Delphi process was not conducted, future studies should be conducted to confirm these findings using larger expert panels and iterative consensus methods. The study was also limited to a specific set of questions pertaining to one phase of DFU management; the findings may not be generalizable to other clinically complex, evidence-scarce domains without further research. Due to a small, highly paired sample and the evaluation of 6 models across 8 countries, only descriptive analysis was conducted. Inferential procedures such as pairwise tests and logistic models can be unstable and prone to error, especially when multiple comparisons are performed, and were therefore used only as a supporting argument for the descriptive analysis. Finally, hallucination was assessed using a binary definition based on fabricated or unverifiable references. A broader grounding assess-

ment, including whether real references were clinically relevant and directly supported the specific LLM statements, was not performed. Future studies should incorporate a more detailed reference-grounding framework distinguishing fabricated sources, irrelevant sources, unsupported claims, and fully supported citations. Moreover, larger cohorts of survey participants, more questions across other uncertain clinical domains, and refined prompts will be required to support these findings.

5. Conclusions

This study assessed 6 major publicly available LLMs in a challenging evidence-scarce clinical scenario on preventive strategies for DFUs. Results suggest that while current LLMs somewhat align with healthcare professionals' opinions, they are unreliable due to limited local context awareness and hallucinations. Most models trained on broad data did not demonstrate improved accuracy for country-specific advice, except for Perplexity and ChatGPT 4-o. Subscription models like ChatGPT 5.0 and Gemini 2.5 Pro were not definitively more reliable than free ones in this context. In the global analysis, model accuracy ranged from 50.0% to 75.0%, while hallucination rates ranged from 12.5% to 100.0%. When accuracy and hallucination were combined into the reliability index, performance ranged from 0% to 54.7%, indicating that models with apparently acceptable accuracy could still be unreliable when their supporting references were invalid. Country-aware prompting did not consistently improve performance, and in several models, was associated with persistent or higher hallucination. These findings suggest that LLMs used at the time of the study might not be suitable as independent clinical decision-support tools during the DFU transition phase. Although the LLMs used in this study are effective at standard medical fact-checking, developing true clinical AI requires addressing practical challenges, such as region-specific adaptation, factual grounding, and expert-level insight generation. Until reliability and contextual awareness are improved, LLMs should serve as tools for human experts, who retain judgment and accountability in patient care.

Author Contributions: Conceptualization, K.S., H.S. and P.C.; methodology, K.S., H.S., G.R. and P.C.; software, K.S.; validation, K.S., H.S., G.R., A.L. and P.C.; formal analysis, K.S. and P.C.; investigation K.S. and H.S.; resources, A.L.; data curation, K.S.; writing—original draft preparation, K.S., H.S. and P.C.; writing—review and editing, K.S., H.S., G.R., A.L., L.B. and P.C.; visualization, K.S.; supervision, G.R., A.L., L.B. and P.C.; project administration, A.L. and L.B.; funding acquisition, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This project received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 101073533 (DIALECT: Diabetes Lower Extremity Complications Research and Training Network in Foot Ulcer and Amputation Prevention).

Informed Consent Statement: Participation in the survey was voluntary, and informed consent was obtained before filling out the questionnaires.

Data Availability Statement: Parsed-JSON data of 192 answers and 384 references, generated by the LLMs, can be provided upon reasonable request from the corresponding author.

Acknowledgments: During the preparation of this work, the authors used Grammarly (version 1.147.1.0) to improve the quality of their English-language expression, tone, and proofreading. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DFU	Diabetic Foot Ulcer
DM	Diabetes Mellitus
AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
LLM	Large Language Model

Appendix A. DIALECT Clinician Survey

The clinical questions were developed by the multidisciplinary author team to cover four pre-specified domains considered relevant to transition-phase offloading: time to long-term footwear, insole characteristics, midsole/outsole characteristics, and upper design. The questionnaire was reviewed internally for clinical relevance and clarity before distribution. The survey used is as follows:

Footwear Solutions for Individuals in the Transition Phase

The following items addressed the standard of care for people with diabetes who had recently recovered from a foot ulcer. This short period, immediately following ulcer closure while tissue properties remain unstable, is referred to as the transition phase.

Demographic Questions

- A. What is your professional background?
- ☐ Podiatrist
 - ☐ Physician
 - ☐ Orthotist/Prosthetist
 - ☐ Shoe Technician
 - ☐ Researcher/Scientist
 - ☐ Other (please specify)
- B. How many years of experience do you have working with diabetic foot patients?
- ☐ <5 years
 - ☐ 5–10 years
 - ☐ 11–20 years
 - ☐ 20 years
- C. In which country do you primarily practice?

Clinical Questions

1. Based on your experience, how long does it typically take for a patient to receive their recommended footwear for long-term use after the ulcer has healed?
 - ☐ Less than 1 month
 - ☐ 1–3 months
 - ☐ 3–6 months
 - ☐ Other (please specify)
2. In your experience, what do patients use during the transition phase, while waiting for their long-term post-ulcer footwear?
 - ☐ Continue with their current footwear
 - ☐ Wear off-the-shelf therapeutic footwear
 - ☐ Wear non-therapeutic footwear
 - ☐ Other (please specify)

3. Do you think special attention is necessary during the *transition phase*, particularly regarding dedicated footwear solutions that meet offloading needs?
 - ☐ Yes
 - ☐ No
 - ☐ I am not sure
4. Which shoe upper type would you recommend for the transition-phase footwear?
 - ☐ Closed (e.g., sneakers)
 - ☐ Open (e.g., sandals)
5. Which insole type would you recommend for the transition-phase footwear?
 - ☐ Custom-made
 - ☐ Semi-custom (customizable)
 - ☐ Off-the-shelf insoles
6. Which midsole/outsole material would you recommend for the transition-phase footwear?
 - ☐ Flexible
 - ☐ Semi-rigid
 - ☐ Rigid
7. Which midsole/outsole shape would you recommend for the transition-phase footwear?
 - ☐ Toe only rocker
 - ☐ Forefoot rocker
 - ☐ Roller design
 - ☐ Half-shoe
8. In your opinion, what improvements could be made to the current standard of care during the transition phase?
 - ☐ (Open-text response)

Appendix B. Promoting Technique

Choice of zero-shot (instruction) prompting.

This study evaluated six LLMs on the same eight clinical questions to benchmark intrinsic capability and contextual awareness. To avoid content leakage and ensure comparability across models and countries, (zero-shot) prompting with a strict JSON schema and role/context conditioning (IWGDF-trained clinician; country-aware conditions) was used. Zero-shot prompting was chosen over few-shot, even though the latter can improve format compliance, because it also risks seeding information or citation style, potentially reducing measured hallucination and altering the construct being benchmarked. Given the study's goal (baseline capability profiling) and the limited item count, unassisted performance was prioritized over exemplar-aided performance.

Global Prompting

First, all the models were given the following general prompt to provide context for the task.

"You are an expert in diabetic foot care and footwear prescription, trained in the International Working Group on the Diabetic Foot (IWGDF) guidelines. Your task is to provide clear, evidence-based recommendations in JSON format. Return ONLY valid JSON. Do not include any text outside the JSON. The questions are designed to collect insights into the current standard of care for people with diabetes who have just recovered from a foot ulcer. This brief period (a few weeks), immediately after the ulcer has healed and while the tissue properties are still changing, is referred to as the 'transition phase'."

Next, all models were asked the same eight questions as in the healthcare professional survey. A few examples are as follows:

Example 1 (Prompt for Q1)

You are answering exactly one option and must reply in strict JSON.

“How long after an ulcer has healed should patients receive their recommended long-term footwear?”

Allowed options (choose exactly one):

- *“Less than 1 month”*
- *“1–3 months”*
- *“3–6 months”*
- *“Other (please specify)”*

Return only this JSON:

```
{
  "answer": "<one of the allowed options>",
  "explanation": "<1–3 lines why this is clinically appropriate>",
  "references": ["<Author Year, Journal>", "<Author Year, Journal>"]
}
```

Example 2 (Prompt for Q2)

You are answering exactly one option and must reply in strict JSON.

“What do patients use during the transition phase, while waiting for their long-term post-ulcer footwear?”

Allowed options (choose exactly one):

- *“Continue with their current footwear”*
- *“Wear off-the-shelf therapeutic footwear”*
- *“Wear non-therapeutic footwear”*
- *“Other (please specify)”*

Return only this JSON:

```
{
  "answer": "<one of the allowed options>",
  "explanation": "<1–3 lines why this is clinically appropriate>",
  "references": ["<Author Year, Journal>", "<Author Year, Journal>"]
}
```

Country-Specific Prompting

After responses to the general prompts for Q1–Q8 were recorded, the models were given the following role-assignment prompt to collect country-aware responses.

The following is the prompt for the Netherlands, similar prompts were used for Denmark and the UK.

“You are an expert in diabetic foot care and footwear prescription, trained in the IWGDF guidelines, and you are aware of local clinical practices, healthcare infrastructure, and reimbursement policies in the Netherlands. Your task is to provide clear, evidence-based recommendations in JSON format. Return ONLY valid JSON. Do not include any text outside the JSON. The questions are designed to collect insights on the current standard of care for people with diabetes who have just recovered from a foot ulcer. This brief period (a few weeks), immediately after the ulcer has healed and while the tissue properties are still changing, is referred to as the ‘transition phase’.”

Following this, all models were asked the same eight questions as in the healthcare professional survey, as shown in Examples 1 and 2 of global prompting.

Appendix C. Question-Wise LLM Performance

Question	Consensus	Model	Answer	Model Accuracy	Hallucination_rate
Q1_time_to_longterm	Consensus = 1–3 months	ChatGPT 4-o	Less than 1 month	0	0
Q1_time_to_longterm	Consensus = 1–3 months	ChatGPT 5.0	Less than 1 month	0	0
Q1_time_to_longterm	Consensus = 1–3 months	Gemini 2.5 Flash	3–6 months	0	1
Q1_time_to_longterm	Consensus = 1–3 months	Gemini 2.5 Pro	1–3 months	1	1
Q1_time_to_longterm	Consensus = 1–3 months	Claude Sonnet 4.0	1–3 months	1	0
Q1_time_to_longterm	Consensus = 1–3 months	Perplexity	1–3 months	1	0
Q2_use_during_transition	Consensus = Other (please specify)	ChatGPT 4-o	Wear off-the-shelf therapeutic footwear	0	1
Q2_use_during_transition	Consensus = Other (please specify)	ChatGPT 5.0	Wear off-the-shelf therapeutic footwear	0	1
Q2_use_during_transition	Consensus = Other (please specify)	Gemini 2.5 Flash	Wear off-the-shelf therapeutic footwear	0	1
Q2_use_during_transition	Consensus = Other (please specify)	Gemini 2.5 Pro	Other (please specify): A transitional healing shoe or the offloading device used to heal the ulcer.	1	0
Q2_use_during_transition	Consensus = Other (please specify)	Claude Sonnet 4.0	Wear off-the-shelf therapeutic footwear	0	1
Q2_use_during_transition	Consensus = Other (please specify)	Perplexity	Wear off-the-shelf therapeutic footwear	0	1
Q3_special_attention	Consensus = Yes	ChatGPT 4-o	Yes	1	1
Q3_special_attention	Consensus = Yes	ChatGPT 5.0	Yes	1	1
Q3_special_attention	Consensus = Yes	Gemini 2.5 Flash	Yes	1	1
Q3_special_attention	Consensus = Yes	Gemini 2.5 Pro	Yes	1	0
Q3_special_attention	Consensus = Yes	Claude Sonnet 4.0	Yes	1	0
Q3_special_attention	Consensus = Yes	Perplexity	Yes	1	1

Question	Consensus	Model	Answer	Model Accuracy	Hallucination_rate
Q4_upper_type	Consensus = Closed (like sneakers)	ChatGPT 4-o	Closed (like sneakers)	1	0
Q4_upper_type	Consensus = Closed (like sneakers)	ChatGPT 5.0	Closed (like sneakers)	1	0
Q4_upper_type	Consensus = Closed (like sneakers)	Gemini 2.5 Flash	Closed (like sneakers)	1	1
Q4_upper_type	Consensus = Closed (like sneakers)	Gemini 2.5 Pro	Closed (like sneakers)	1	1
Q4_upper_type	Consensus = Closed (like sneakers)	Claude Sonnet 4.0	Closed (like sneakers)	1	0
Q4_upper_type	Consensus = Closed (like sneakers)	Perplexity	Closed (like sneakers)	1	1
Q5_insole_type	Consensus = Custom-made	ChatGPT 4-o	Semi-custom (customizable)	0	0
Q5_insole_type	Consensus = Custom-made	ChatGPT 5.0	Custom-made	1	1
Q5_insole_type	Consensus = Custom-made	Gemini 2.5 Flash	Semi-custom (customizable)	0	1
Q5_insole_type	Consensus = Custom-made	Gemini 2.5 Pro	Custom-made	1	0
Q5_insole_type	Consensus = Custom-made	Claude Sonnet 4.0	Semi-custom (customizable)	0	0
Q5_insole_type	Consensus = Custom-made	Perplexity	Semi-custom (customizable)	0	1
Q6_material_properties	Consensus = Semi-rigid	ChatGPT 4-o	Semi-rigid	1	1
Q6_material_properties	Consensus = Semi-rigid	ChatGPT 5.0	Semi-rigid	1	1
Q6_material_properties	Consensus = Semi-rigid	Gemini 2.5 Flash	Semi-rigid	1	1
Q6_material_properties	Consensus = Semi-rigid	Gemini 2.5 Pro	Rigid	0	1
Q6_material_properties	Consensus = Semi-rigid	Claude Sonnet 4.0	Semi-rigid	1	0
Q6_material_properties	Consensus = Semi-rigid	Perplexity	Semi-rigid	1	1
Q7_shape	Consensus = Forefoot rocker	ChatGPT 4-o	Forefoot rocker	1	0
Q7_shape	Consensus = Forefoot rocker	ChatGPT 5.0	Forefoot rocker	1	1
Q7_shape	Consensus = Forefoot rocker	Gemini 2.5 Flash	Forefoot rocker	1	1

Question	Consensus	Model	Answer	Model Accuracy	Hallucination_rate
Q7_shape	Consensus = Forefoot rocker	Gemini 2.5 Pro	Forefoot rocker	1	0
Q7_shape	Consensus = Forefoot rocker	Claude Sonnet 4.0	Forefoot rocker	1	0
Q7_shape	Consensus = Forefoot rocker	Perplexity	Forefoot rocker	1	1
Q8_Improvements	Consensus = Availability/Variety	ChatGPT 4-o	Introduce standardized protocols for timely provision of interim therapeutic footwear and monitor adherence through structured follow-up.	0	0
Q8_Improvements	Consensus = Availability/Variety	ChatGPT 5.0	Developing dedicated transition-phase footwear that combines immediate availability, protective uppers, custom insoles, and rocker soles could improve continuity of care and reduce early re-ulceration risk.	1	0
Q8_Improvements	Consensus = Availability/Variety	Gemini 2.5 Flash	There is a need for clearer, evidence-based guidelines specifically for the transition phase after diabetic foot ulcer healing. The current IWGDF guidelines address footwear for primary and secondary prevention but are less specific on the immediate post-ulcer period. This leads to variability in practice and may compromise patient safety.	0	1
Q8_Improvements	Consensus = Availability/Variety	Gemini 2.5 Pro	The greatest improvement would be the integration of smart technology, such as in-shoe pressure and temperature monitoring systems. These wearables provide real-time feedback to patients and clinicians, allowing for the early detection of high-risk activity and inflammation before re-ulceration occurs.	0	1

Question	Consensus	Model	Answer	Model Accuracy	Hallucination_rate
Q8_Improvements	Consensus = Availability/Variety	Claude Sonnet 4.0	Standardized transition protocols with defined timelines, validated pressure monitoring systems for real-time offloading assessment, and structured patient education programs with adherence monitoring would significantly improve outcomes. Implementation of telemedicine surveillance for early complication detection and mandatory multidisciplinary team coordination between podiatrists, orthotists, and endocrinologists during this critical period.	0	0
Q8_Improvements	Consensus = Availability/Variety	Perplexity	Incorporate digital monitoring, multidisciplinary follow-up, and timely access to customizable offloading solutions.	0	0

References

1. Bhuyan, S.S.; Sateesh, V.; Mukul, N.; Galvankar, A.; Mahmood, A.; Nauman, M.; Rai, A.; Bordoloi, K.; Basu, U.; Samuel, J. Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *J. Med. Syst.* **2025**, *49*, 10. [[CrossRef](#)] [[PubMed](#)]
2. Lin, C.; Kuo, C.-F. Roles and Potential of Large Language Models in Healthcare: A Comprehensive Review. *Biomed. J.* **2025**, *48*, 100868. [[CrossRef](#)] [[PubMed](#)]
3. Artsi, Y.; Sorin, V.; Glicksberg, B.S.; Korfiatis, P.; Freeman, R.; Nadkarni, G.N.; Klang, E. Challenges of Implementing LLMs in Clinical Practice: Perspectives. *J. Clin. Med.* **2025**, *14*, 6169. [[CrossRef](#)] [[PubMed](#)]
4. Hossain, M.J.; Al-Mamun, M.; Islam, M.R. Diabetes Mellitus, the Fastest Growing Global Public Health Concern: Early Detection Should Be Focused. *Health Sci. Rep.* **2024**, *7*, e2004. [[CrossRef](#)] [[PubMed](#)]
5. Xu, T.; Hu, L.; Xie, B.; Huang, G.; Yu, X.; Mo, F.; Li, W.; Zhu, M. Analysis of Clinical Characteristics in Patients with Diabetic Foot Ulcers Undergoing Amputation and Establishment of a Nomogram Prediction Model. *Sci. Rep.* **2024**, *14*, 27934. [[CrossRef](#)] [[PubMed](#)]
6. Armstrong, D.G.; Boulton, A.J.M.; Bus, S.A. Diabetic Foot Ulcers and Their Recurrence. *N. Engl. J. Med.* **2017**, *376*, 2367–2375. [[CrossRef](#)] [[PubMed](#)]
7. Lazzarini, P.A.; van Netten, J.J. Best Practice Offloading Treatments for Diabetic Foot Ulcer Healing, Remission, and Better Plans for the Healing-Remission Transition. *Semin. Vasc. Surg.* **2025**, *38*, 110–120. [[CrossRef](#)] [[PubMed](#)]
8. Bus, S.A.; Armstrong, D.G.; Crews, R.T.; Gooday, C.; Jarl, G.; Kirketerp-Moller, K.; Viswanathan, V.; Lazzarini, P.A. Guidelines on Offloading Foot Ulcers in Persons with Diabetes (IWGDF 2023 Update). *Diabetes/Metab. Res. Rev.* **2024**, *40*, e3647. [[CrossRef](#)] [[PubMed](#)]
9. Bus, S.A.; Sacco, I.C.N.; Monteiro-Soares, M.; Raspovic, A.; Paton, J.; Rasmussen, A.; Lavery, L.A.; Netten, J.J. van Guidelines on the Prevention of Foot Ulcers in Persons with Diabetes (IWGDF 2023 Update). *Diabetes/Metab. Res. Rev.* **2024**, *40*, e3651. [[CrossRef](#)] [[PubMed](#)]
10. Sarlak, H.; Shakir, K.; Rogati, G.; Leardini, A.; Berti, L.; Caravaggi, P. Current Strategies and Priorities for Diabetic Footwear Design and Production: A Cross-European Exploratory Survey of Clinicians and Shoemakers. *Acta Diabetol.* **2026**, *63*, 887–895. [[CrossRef](#)] [[PubMed](#)]
11. Jones, A.W.; Makanjuola, A.; Bray, N.; Prior, Y.; Parker, D.; Nester, C.; Tang, J.; Jiang, L. The Efficacy of Custom-Made Offloading Devices for Diabetic Foot Ulcer Prevention: A Systematic Review. *Diabetol. Metab. Syndr.* **2024**, *16*, 172. [[CrossRef](#)] [[PubMed](#)]
12. Shakir, K.; Sarlak, H.; Bus, S.A.; Rogati, G.; Leardini, A.; Berti, L.; Caravaggi, P. Advancing Diabetic Foot Ulcer Care: Focus on the Post-Healing Transition Phase. *J. Diabetes* **2025**, *17*, e70173. [[CrossRef](#)] [[PubMed](#)]
13. Schaper, N.C.; Van Netten, J.J.; Apelqvist, J.; Bus, S.A.; Fitridge, R.; Game, F.; Monteiro-Soares, M.; Senneville, E.; The IWGDF Editorial Board. Practical Guidelines on the Prevention and Management of Diabetes-related Foot Disease (IWGDF 2023 Update). *Diabetes Metab. Res.* **2024**, *40*, e3657. [[CrossRef](#)] [[PubMed](#)]
14. Tenny, S.; Varacallo, M.A. Evidence-Based Medicine. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2025.
15. Burns, P.B.; Rohrich, R.J.; Chung, K.C. The Levels of Evidence and Their Role in Evidence-Based Medicine. *Plast. Reconstr. Surg.* **2011**, *128*, 305–310. [[CrossRef](#)] [[PubMed](#)]
16. Mashatian, S.; Armstrong, D.G.; Ritter, A.; Robbins, J.; Aziz, S.; Alenabi, I.; Huo, M.; Anand, T.; Tavakolian, K. Building Trustworthy Generative Artificial Intelligence for Diabetes Care and Limb Preservation: A Medical Knowledge Extraction Case. *J. Diabetes Sci. Technol.* **2025**, *19*, 1264–1270. [[CrossRef](#)] [[PubMed](#)]
17. Rohrich, R.N.; Li, K.R.; Lava, C.X.; Snee, I.; Alahmadi, S.; Youn, R.C.; Steinberg, J.S.; Atves, J.M.; Attinger, C.E.; Evans, K.K. Consulting the Digital Doctor: Efficacy of ChatGPT-3.5 in Answering Questions Related to Diabetic Foot Ulcer Care. *Adv. Skin Wound Care* **2025**, *38*, E74–E80. [[CrossRef](#)] [[PubMed](#)]
18. Popa, A.D.; Gavril, R.S.; Popa, I.V.; Mihalache, L.; Gherasim, A.; Niță, G.; Graur, M.; Arhire, L.I.; Niță, O. Survival Prediction in Diabetic Foot Ulcers: A Machine Learning Approach. *J. Clin. Med.* **2023**, *12*, 5816. [[CrossRef](#)] [[PubMed](#)]
19. Misir, A. Artificial Intelligence and Machine Learning in Diabetic Foot Ulcer Care: Advances in Diagnosis, Treatment, Prognosis, and Novel Therapeutic Strategies. *J. Diabetes Sci. Technol.* **2025**, *in press*. [[CrossRef](#)] [[PubMed](#)]
20. Shiraishi, M.; Lee, H.; Kanayama, K.; Moriwaki, Y.; Okazaki, M. Appropriateness of Artificial Intelligence Chatbots in Diabetic Foot Ulcer Management. *Int. J. Low. Extrem. Wounds* **2024**, *25*, 462–468. [[CrossRef](#)] [[PubMed](#)]
21. Gumilar, K.E.; Indraprasta, B.R.; Hsu, Y.-C.; Yu, Z.-Y.; Chen, H.; Irawan, B.; Tambunan, Z.; Wibowo, B.M.; Nugroho, H.; Tjokroprawiro, B.A.; et al. Disparities in Medical Recommendations from AI-Based Chatbots across Different Countries/Regions. *Sci. Rep.* **2024**, *14*, 17052. [[CrossRef](#)] [[PubMed](#)]
22. Yu, E.; Chu, X.; Zhang, W.; Meng, X.; Yang, Y.; Ji, X.; Wu, C. Large Language Models in Medicine: Applications, Challenges, and Future Directions. *Int. J. Med. Sci.* **2025**, *22*, 2792–2801. [[CrossRef](#)] [[PubMed](#)]
23. Greenhalgh, T.; Howick, J.; Maskrey, N. For the Evidence Based Medicine Renaissance Group Evidence Based Medicine: A Movement in Crisis? *BMJ* **2014**, *348*, g3725. [[CrossRef](#)] [[PubMed](#)]

24. Littlejohns, P.; Cluzeau, F.; Thomason, M. Guideline Development in Europe: An International Comparison. *Int. J. Technol. Assess. Health Care* **2000**, *16*, 1039–1049. [[CrossRef](#)] [[PubMed](#)]
25. Ortopædisk Fodtøj og Indlæg. Available online: <https://www.retsinformation.dk/eli/lta/2017/1247> (accessed on 12 October 2025).
26. Caravaggi, P.; Giangrande, A.; Lullini, G.; Padula, G.; Berti, L.; Leardini, A. In Shoe Pressure Measurements during Different Motor Tasks While Wearing Safety Shoes: The Effect of Custom Made Insoles vs. Prefabricated and off-the-Shelf. *Gait Posture* **2016**, *50*, 232–238. [[CrossRef](#)] [[PubMed](#)]
27. Wada, A.; Tanaka, Y.; Nishizawa, M.; Yamamoto, A.; Akashi, T.; Hagiwara, A.; Hayakawa, Y.; Kikuta, J.; Shimoji, K.; Sano, K.; et al. Retrieval-Augmented Generation Elevates Local LLM Quality in Radiology Contrast Media Consultation. *npj Digit. Med.* **2025**, *8*, 395. [[CrossRef](#)] [[PubMed](#)]
28. Wu, K.; Wu, E.; Wei, K.; Zhang, A.; Casasola, A.; Nguyen, T.; Riantawan, S.; Shi, P.; Ho, D.; Zou, J. An Automated Framework for Assessing How Well LLMs Cite Relevant Medical References. *Nat. Commun.* **2025**, *16*, 3615. [[CrossRef](#)] [[PubMed](#)]
29. Shekar, S.; Pataranutaporn, P.; Sarabu, C.; Cecchi, G.A.; Maes, P. People Overtrust AI-Generated Medical Advice despite Low Accuracy. *NEJM AI* **2025**, *2*, A10a2300015. [[CrossRef](#)]
30. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [[CrossRef](#)] [[PubMed](#)]
31. Abdelwanis, M.; Alarafati, H.K.; Tammam, M.M.S.; Simsekler, M.C.E. Exploring the Risks of Automation Bias in Healthcare Artificial Intelligence Applications: A Bowtie Analysis. *J. Saf. Sci. Resil.* **2024**, *5*, 460–469. [[CrossRef](#)]
32. Esmaeilzadeh, P. Ethical Implications of Using General-Purpose LLMs in Clinical Settings: A Comparative Analysis of Prompt Engineering Strategies and Their Impact on Patient Safety. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 342. [[CrossRef](#)] [[PubMed](#)]
33. Luo, M.; Yousefirizi, F.; Rouzrokh, P.; Jin, W.; Alberts, I.; Gowdy, C.; Bouchareb, Y.; Hamarneh, G.; Klyuzhin, I.; Rahmim, A. Physician-in-the-Loop Active Learning in Radiology Artificial Intelligence Workflows: Opportunities, Challenges, and Future Directions. *Am. J. Roentgenol.* **2025**, *225*, AJR.25.33364. [[CrossRef](#)] [[PubMed](#)]
34. Yi, P.H.; Haver, H.L.; Jeudy, J.J.; Kim, W.; Kitamura, F.C.; Oluyemi, E.T.; Smith, A.D.; Moy, L.; Parekh, V.S. Best Practices for the Safe Use of Large Language Models and Other Generative AI in Radiology. *Radiology* **2025**, *316*, e241516. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.