Allergologia et immunopathologia
Sociedad Española de Inmunología Clínica, Alergología y Asma Pediátrica
www.all-imm.com

ORIGINAL ARTICLE

OPEN ACCESS

Evaluating large language models for WAO/EAACI guideline compliance in hereditary angioedema management

Mehmet Emin Gerek*, Tuğba Önalan, Fatih Çölkesen, Şevket Arslan

Department of Internal Medicine, Division of Clinical Immunology and Allergy, Necmettin Erbakan University, Faculty of Medicine, Konya, Türkiye

Received 5 March 2025; Accepted 21 May 2025
Available online 1 July 2025

KEYWORDS
artificial Intelligence;
hereditary angioedema;
large language models;
medical guidelines compliance;
WAO/EAACI guidelines

Abstract
Introduction: Hereditary angioedema (HAE) is a rare but potentially life-threatening disorder characterized by recurrent swelling episodes. Adherence to clinical guidelines, such as the World Allergy Organization/European Academy of Allergy & Clinical Immunology (WAO/EAACI) guidelines, is crucial for effective management. With the increasing role of artificial intelligence in medicine, large language models (LLMs) offer potential for clinical decision support. This study evaluates the performance of ChatGPT, Gemini, Perplexity, and Copilot in providing guideline-adherent responses for HAE management.
Methods: Twenty-eight key recommendations from the WAO/EAACI HAE guidelines were reformulated into interrogative formats and posed to the selected LLMs. Two independent clinicians assessed responses based on accuracy, adequacy, clarity, and citation reliability using a five-point Likert scale. References were categorized as guideline-based, trustworthy, or untrustworthy. A reevaluation with explicit citation instructions was conducted, with discrepancies resolved by a third reviewer.
Results: ChatGPT and Gemini outperformed Perplexity and Copilot, achieving median accuracy and adequacy scores of 5.0 versus 3.0, respectively. ChatGPT had the lowest rate of unreliable references, whereas Gemini showed inconsistency in citation behavior. Significant differences in response quality were observed among models ($p < 0.001$). Providing explicit sourcing instructions improved performance consistency, particularly for Gemini.
Conclusion: ChatGPT and Gemini demonstrated superior adherence to WAO/EAACI guidelines, suggesting that LLMs can support clinical decision-making in rare diseases. However, inconsistencies in citation practices highlight the need for further validation and optimization to enhance reliability in medical applications.

***Corresponding author**: Mehmet Emin Gerek, Necmettin Erbakan University, Faculty of Medicine, Department of Internal Medicine, Division of Clinical Immunology and Allergy, Konya, Türkiye. *Email address*: drmegerek@gmail.com

https://doi.org/10.15586/aei.v53i4.1353

Introduction

Integrating artificial intelligence (AI) into healthcare has markedly improved the accessibility, precision, and efficiency of medical decision-making. Among recent advances, large language models (LLMs) such as ChatGPT, Perplexity, Gemini, and Copilot have garnered considerable attention for their potential in clinical applications, ranging from patient education to treatment planning.[1,2] By bridging extensive medical knowledge with practical clinical insights, these models show promise for addressing rare and complex medical conditions.

AI chatbots have demonstrated significant proficiency in providing individualized patient information. Research indicates that LLMs are capable of tackling inquiries concerning disease management, including keratoconus and urolithiasis while adhering to recognized clinical guidelines across multiple disciplines.[3,4] ChatGPT has consistently demonstrated accuracy in line with validated guidelines, achieving correct or highly accurate responses in up to 97% of instances when evaluated against the recommendations of the European Association of Urology (EAU).[3] Nonetheless, inconsistencies in reference attribution and sporadic dependence on unverified sources continue to pose substantial challenges.[1,5]

Although LLMs are widely used, they differ markedly in architecture, data sources, and response generation strategies. For instance, ChatGPT relies on an extensive training dataset that includes peer-reviewed medical literature but lacks real-time internet access, limiting its ability to provide up-to-date citations. Gemini, developed by Google, integrates real-time web searches but has been noted for inconsistent citation practices. Perplexity AI emphasizes reference-based responses and transparency in its citations, making it particularly relevant for clinical guideline adherence. In contrast, Copilot, developed by Microsoft, is integrated into productivity software and designed for general information retrieval rather than specialized medical queries. These inherent differences contribute to variability in their performance across different medical domains, including hereditary angioedema (HAE) management.

HAE is a rare, potentially fatal condition marked by recurrent episodes of significant swelling in multiple regions of the body. Effective management of this condition relies on compliance with established protocols, such as the International WAO/EAACI Guideline for the Management of Hereditary Angioedema—The 2021 Revision and Update. These guidelines offer a systematic diagnosis, treatment, and long-term care framework, highlighting evidence-based methodologies to enhance patient outcomes.[2,3]

Despite the availability of these evidence-based recommendations, real-world adherence remains suboptimal, particularly in centers with limited HAE expertise or emergency settings where rapid, guideline-concordant decisions are critical.[6] Moreover, a recent multinational survey reported that fewer than half of the physicians managing rare angioedema cases feel fully confident in applying the 2021 World Allergy Organization/European Academy of Allergy & Clinical Immunology (WAO/EAACI) update, citing time constraints and unfamiliarity with specialized treatments as the key barriers.[7] Scalable digital tools capable of instantly translating guideline knowledge into clear clinical answers could therefore narrow the expertise gap, reduce diagnostic delays, and ultimately improve patient outcomes. However, no study has systematically examined whether state-of-the-art LLMs can deliver HAE recommendations that mirror WAO/EAACI guidance. Addressing this question is pivotal before such models can be considered for decision support or patient-facing applications.

Although guideline adherence is of paramount importance, further research is required to evaluate LLMs' efficacy in implementing condition-specific recommendations, such as those pertaining to HAE. Prior assessments have underscored the capacity of AI to assist healthcare practitioners in making evidence-based decisions. A study assessing LLMs in breast cancer screening indicated that they could equal or surpass the diagnostic accuracy of radiologists in identifying malignancies, highlighting their potential in clinical practice.[8] However, the study also highlighted the inconsistency in model performance, stressing the necessity for thorough validation.[9]

Moreover, reference attribution is crucial in assessing the reliability of LLM-generated information. Previous studies on the efficacy of LLM-based chatbots in medical educational resources for conditions such as epilepsy and cardiac catheterization have demonstrated disparate levels of reliability in the sources referenced by these models.[10,11] The findings indicate that although LLM-powered chatbots offer potential across various fields, their dependence on unverified sources may compromise their credibility and relevance in clinical decision-making.

This study seeks to assess the capacity of ChatGPT-4o, Gemini 2.0, Perplexity, and Copilot to deliver responses that align with the WAO/EAACI guidelines for HAE management.[12] This research also aimed to address essential inquiries regarding the clinical applicability of LLMs in specialized medical fields by transforming 28 guideline recommendations into interrogative formats and systematically evaluating the accuracy, adequacy, and clarity of their responses. The findings are anticipated to clarify the advantages and drawbacks of LLMs, enhancing the overarching discourse on their function in evidence-based medicine and patient care.

Materials and Methods

This study evaluated the adherence of four LLMs—ChatGPT-4o, Gemini 2.0, Perplexity, and Copilot—to the International WAO/EAACI Guideline for the Management of Hereditary Angioedema—The 2021 Revision and Update. A structured methodology assessed LLM-generated responses, accuracy, adequacy, clarity, and citation reliability.

Question formulation

All 28 evidence-based recommendations from the WAO/EAACI guideline were converted into closed-ended interrogative formats (yes/no or specific-choice questions) to strictly evaluate guideline adherence. The transformation adhered to the following protocol.

Direct rephrasing

Recommendations were rephrased verbatim without altering their clinical intent.

For example:

- Guideline: *We recommend C1 [a protein found in the fluid part of the blood] inhibitor or icatibant be used for the treatment of attacks in children under 12.*
- Question: *Is it recommended to use C1 inhibitor or icatibant for the treatment of attacks in HAE-diagnosed children under the age of 12?*

Full question list:

The complete set of 28 questions is provided in the supplementary file.

Model interaction protocol

Prompt design:

Questions were posed directly without additional context (e.g., no guideline-specific prompts).

For example:

- *Should all patients suspected of having HAE be assessed for blood levels of [C1 esterase inhibitor]C1-INH function, C1-INH protein, and C4?*

Model parameters:

All LLMs were used with their default settings to ensure consistency in response generation. No adjustments were made to parameters such as temperature or token limits. Model versions included ChatGPT-4o (latest as of December 2024), Gemini 2.0 (December 2024), Perplexity (December 2024), and Copilot (Bing, December 2024).

- Session management:
- Prior interactions were cleared before each query to prevent bias from previous conversations.

Evaluation process

Blinded assessment by clinicians

Two independent clinicians with expertise in HAE management evaluated the responses. To ensure objectivity, observers were blinded to the LLM source of each response and evaluators were active in HAE diagnosis, treatment, and guideline implementation in clinical practice.

Each response was evaluated based on the following four criteria:

- Accuracy: How well does the response align with the guideline?
- Adequacy: Whether the response provides sufficient information to address the question.
- Clarity: The level of comprehensibility and coherence of the response.
- Citation Use: Whether the response includes a reference and the trustworthiness of the cited source.

A five-point Likert scale was utilized to evaluate responses based on their accuracy, adequacy, clarity, and citation reliability, in accordance with the International WAO/EAACI Guideline for the Management of Hereditary Angioedema—The 2021 Revision and Update. The scale was defined as:[3]

1. Incorrect: The response contradicts the International WAO/EAACI Guideline or contains factual inaccuracies.
2. Insufficient: The response is partially correct but lacks critical details or omits essential information.
3. Sufficient: The response demonstrates basic alignment with the International WAO/EAACI Guideline, yet lacks specificity or depth.
4. Accurate: The response adheres to the International WAO/EAACI Guideline, but lacks comprehensive detail or nuanced interpretation.
5. Very Accurate: The response fully complies with the International WAO/EAACI Guideline, providing detailed, precise, and guideline-consistent information.

Third reviewer for discrepancies

In cases of disagreement (> 1-point score difference), an independent HAE specialist with peer-reviewed publications in the field resolved all conflicts. This reviewer had extensive experience in HAE research and clinical care, ensuring consensus-based final scores.

Reference classification

Reference reliability was a critical aspect of the evaluation process. The LLMs were explicitly instructed to prioritize peer-reviewed and evidence-based references to ensure a rigorous assessment of citation quality. Each question was also evaluated for reference attribution, and references were classified into three categories:

Category 1: References to The International WAO/EAACI Guideline for the Management of Hereditary Angioedema—The 2021 Revision and Update.

Category 2: References to trustworthy sources, such as peer-reviewed scientific publications and the websites of international disease associations managed by clinicians specializing in HAE, referencing peer-reviewed publications.

Category 3: References to untrustworthy sources, including nonpeer-reviewed publications, patient communication communities, websites not managed by specialized clinicians or those lacking references to peer-reviewed materials, and promotional content lacking independent validation.

Re-evaluation of contextual information

To examine the effect of contextual prompting, the entire 28-question evaluation was repeated, informing the LLMs of the user's medical background and instructing them to use trustworthy sources. Blinded assessment protocols were maintained; the same clinicians scored both responses and discrepancies were again resolved by the same independent HAE specialist.

Ensuring impartiality

To minimize bias, each LLM's history was cleared before every session, and observers were blinded to the source of each response. Standardized prompts were applied uniformly.

Each model was tested independently to avoid cross-contamination, aligning with best practices in AI evaluation and enhancing the validity of between-model comparisons.

Statistical analysis

Data was analyzed using IBM SPSS (Statistical Package for the Social Sciences) Statistics version 22.0. Descriptive statistics are presented as frequencies (n) and percentages (%) for categorical variables and as medians and interquartile ranges (IQR) for numerical variables. The Pearson chi-square and Fisher's exact tests were used for categorical comparisons, with Dunn-Bonferroni adjustments for post hoc analyses. Normality of numerical variables was assessed with the Shapiro-Wilk test. Non-normally distributed data were analyzed using nonparametric tests (Mann-Whitney U, Kruskal-Wallis), with pairwise Dunn-Bonferroni corrections for multiple comparisons. Cohen's kappa statistic measured interobserver agreement, interpreted as < 0 (poor), 0.01–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect). Statistical significance was set at $p < 0.05$.

Results

Interrater and RATER model agreement analysis

Cohen's Kappa analysis revealed variability in agreement levels across observers and models (Table 1). The highest interobserver agreement was observed for ChatGPT-1 (κ = 0.418, moderate), whereas ChatGPT-2 achieved the strongest agreement between the first observer and the rater (κ = 0.795, substantial). Perplexity-2 showed the highest concordance between the second observer and the rater (κ = 0.761, substantial). These results underscore that while interobserver agreement was generally moderate, RATER (reliability, assurance, tangibles, empathy, and responsiveness) model agreement varied significantly, with ChatGPT and Perplexity demonstrating the most consistency.

Comparison of first and second responses of LLMs

A significant improvement was observed only for Gemini 2.0 between its first and second responses (Wilcoxon Z = −2.539, $p = 0.011$) (Table 2). In contrast, ChatGPT, Perplexity, and Copilot showed no statistically significant changes. This suggests that contextual prompting (e.g., specifying the user's medical background) selectively enhanced Gemini's adherence to guidelines.

Comparison between LLMs for first and second responses

First response comparison

Kruskal-Wallis analysis indicated significant disparities in initial responses ($\chi^2 = 30.012$, $p < 0.001$). ChatGPT and Gemini outperformed Perplexity and Copilot, achieving median accuracy and adequacy scores of 5.0 versus 3.0, respectively. Post hoc tests revealed that Copilot scored significantly lower than ChatGPT and Perplexity ($p < 0.001$), whereas the latter two performed comparably.

Table 1 Cohen's kappa analysis of inter-observer and observer-rater agreement.

	Observer 1—Observer 2	Observer 1—Rater	Observer 2—Rater
ChatGPT-1st	0.418	0.702	0.524
ChatGPT-2nd	0.278	0.795	0.489
Perplexity-1st	0.300	0.537	0.067
Perplexity-2nd	0.080	0.272	0.761
Gemini-1st	0.128	0.653	0.203
Gemini-2nd	0.397	0.691	0.608
Copilot-1st	0.225	0.417	0.694
Copilot-2nd	0.359	0.632	0.661

Kappa values represent the level of agreement: 0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect).

Table 2 Comparison of first- and second-round responses of large language models (LLMs) and their Intercomparisons.

	First Response		Second Response		z	p
	Median (IQR)	Mean Rank	Median (IQR)	Mean Rank		
ChatGPT[a]	5.0 (4.0-5.0)	69.77	5.0 (4.0-5.0)	62.84	0.957	0.338
Perplexity[b]	5.0 (4.0-5.0)	70.32	5.0 (4.0-5.0)	65.59	0.378	0.705
Gemini[c]	5.0 (1.75-5.0)	54.27	5.0 (5.0-5.0)	69.16	2.539	**0.011**
Copilot[d]	3.0 (2.0-4.0)	31.64	3.0 (2.0-4.0)	28.41	0.773	0.773
	$X^2 = 30.0123$ *p < 0.001 p(a-d) = < 0.001 p(b-d) = < 0.001		$X^2 = 34.6709$ *p < 0.001 p(a-d) = < 0.001 p(b-d) = < 0.001 p(c-d) = < 0.001			

a: ChatGPT; b: Perplexity; c: Gemini; d: Copilot.
Z: Wilcoxon Test Value.
x^2: Kruskal-Wallis Test.
*: Mann-Whitney U Test.

Second response comparison

For the second evaluation, Copilot remained the lowest performer (χ^2 = 34.671, p < 0.001), whereas Gemini 2.0 showed marked improvement, aligning more closely with WAO/EAACI guidelines (p < 0.001).

Reference attribution analysis

Overall reference rates

No significant changes were observed between attempts for individual models. However, in the first attempt, ChatGPT-4o and Gemini 2.0 cited fewer references than Perplexity and Copilot (χ^2 = 98.215, p < 0.001). In the second attempt, Gemini 2.0 referenced fewer sources than all other models (p < 0.001), highlighting inconsistent citation behavior.

Unreliable references

ChatGPT-4o consistently provided the most reliable references (0% unreliable citations in both attempts), whereas Perplexity and Copilot frequently cited unverified sources (50–53.6% unreliable citations, p < 0.001) (Table 3).

EAACI-compliant versus noncompliant references

Compliance with guidelines

Perplexity and Copilot prioritized EAACI-compliant references (78.6–100% compliance), whereas ChatGPT-4o and Gemini 2.0 exhibited variability (p < 0.001) (Table 4).

Noncompliant references

ChatGPT-4o's second attempt showed a surge in noncompliant references (71.4% noncompliant vs. 0% initially,

Table 3 Comparison of references cited by LLMs in the first- and second-round responses.

	References			Unreliable References		
	First	Second	p	First	Second	p
ChatGPT[a]	0 (0.0)	20 (71.4)	-	0 (0.0)	0 (0.0)	-
Perplexity[b]	28 (100.0)	28 (100.0)	-	14 (50.0)	15 (53.6)	1.000*
Gemini[c]	4 (14.3)	8 (28.6)	0.344**	0 (0.0)	5 (17.9)	0.063**
Copilot[d]	28 (100.0)	28 (100.0)	-	13 (46.4)	14 (50.0)	1.000*
	χ^2 = 98.215	χ^2 = 51.048		χ^2 = 35.674	χ^2 = 26.522	
	*p < 0.001	*p < 0.001		*p < 0.001	*p < 0.001	
	**p(a-b) < 0.001	*p(a-b) = 0.024		**p(a-b) < 0.001	**p (a-b) < 0.001	
	**p(a-d) < 0.001	*p (a-c) = 0.018		**p(a-d) < 0.001	**p(a-d) < 0.001	
	**p(b-c) < 0.001	*p(a-d) = 0.024		**p(b-c) < 0.001		
	**p(c-d) < 0.001	*p(b-c) < 0.001		**p(c-d) < 0.001		
		*p(c-d) < 0.001				

a: ChatGPT; b: Perplexity; c: Gemini; d: Copilot.
X²: Chi-Square Test Statistic *: Pearson Chi-Square Test **: Fisher's Exact Test.

Table 4 EAACI-compliant versus noncompliant references across first- and second-round responses.

	WAO/EAACI-Compliant References			WAO/EAACI Noncompliant Reliable References		
	First	Second	p	First	Second	p
ChatGPT[a]	0 (0.0)	0 (0.0)	-	0 (0.0)	20 (71.4)	< 0.001**
Perplexity[b]	22 (78.6)	18 (64.3)	0.344*	28 (100.0)	28 (100.0)	-
Gemini[c]	3 (10.7)	3 (10.7)	1.000**	2 (7.1)	4 (14.3)	0.625**
Copilot[d]	22 (78.6)	20 (71.4)	0.625*	19 (67.9)	20 (71.4)	1.000*
	χ^2 = 62.287	χ^2 = 48.132		χ^2 = 79.637	χ^2 = 47.289	
	*p < 0.001	*p < 0.001		*p < 0.001	*p < 0.001	
	**p(a-b) < 0.001	**p(a-b) < 0.001		**p(a-b) < 0.001	*p(a-b) = 0.024	
	**p(a-d) < 0.001	**p(a-d) < 0.001		**p(a-d) < 0.001	**p(a-c) < 0.001	
	** p(b-c) < 0.001	**p(b-c) < 0.001		**p(b-c) < 0.001	**p(b-c) < 0.001	
	**p(c-d) < 0.001	**p(c-d) < 0.001		*p(b-d) = 0.012	*p(b-d) = 0.024	
				**p(c-d) < 0.001	**p(c-d) < 0.001	

a: ChatGPT; b: Perplexity; c: Gemini; d: Copilot.
*: Pearson Chi-Square Test **: Fisher's Exact Test.

$p < 0.001$), whereas Gemini 2.0 maintained lower rates compared to others ($p < 0.001$).

Correlation between agreement and response quality

Spearman correlation analysis (Table 5) revealed distinct patterns in the relationship between interrater agreement and response accuracy. For the first responses, Gemini 2.0 exhibited a positive correlation between accuracy and observer agreement ($\rho = 0.503$, $p = 0.006$), suggesting its performance improved with higher consensus among evaluators. In the second evaluation round, all models demonstrated strong correlations between agreement levels and response scores, with ChatGPT-4o ($\rho = 0.573$, $p = 0.001$) and Copilot ($\rho = 0.498$, $p = 0.007$) showing particularly robust associations. These findings underscore the critical role of observer consensus in enhancing the reliability of LLM-generated responses, particularly when contextual instructions are provided.

Discussion

This study evaluated four LLMs—ChatGPT-4o, Gemini 2.0, Copilot, and Perplexity—for their adherence to the WAO/EAACI guidelines in HAE management. The results indicate that ChatGPT-4o and Gemini 2.0 generally provided more guideline-concordant and clinically actionable recommendations, whereas Copilot and Perplexity often produced incomplete or inconsistent responses. These observations are compatible with other investigations suggesting that specific LLMs can offer high compliance with medical guidelines; however, performance varies considerably based on the model's training data and disease-specific expertise.[13,14] One publication on resuscitation protocols described a similar finding, noting that some models delivered accurate information, whereas, in contrast, others generated potentially harmful advice, underscoring the importance of verifying AI outputs before clinical use.[15]

These findings are clinically meaningful for at least two reasons. First, HAE is a rare disorder in which many frontline providers lack day-to-day experience; even modest improvements in guideline adherence can therefore translate into tangible reductions in diagnostic delay and attack-related morbidity. Second, to our knowledge, this is the first head-to-head assessment of multiple cutting-edge LLMs against the 2021 WAO/EAACI recommendations, providing an empirical benchmark for future model development. By demonstrating both the strengths and the current shortcomings of LLM outputs, our study offers a roadmap for how AI tools might be fine-tuned or regulated before they can be safely integrated into specialist allergy-immunology workflows.

In most of the questions posed here, ChatGPT-4o and Gemini 2.0 accurately identified first-line therapies such as C1 inhibitor or icatibant for pediatric HAE attacks, demonstrating a strong capacity to interpret nuanced guideline recommendations.[12] In contrast, Copilot and Perplexity frequently omitted prophylaxis protocols, highlighting that general-purpose models may underperform in rare or complex diseases. A study examining older versions of ChatGPT in oncology similarly reported that complicated treatment algorithms exposed the model's limitations, emphasizing the iterative nature of AI improvement and the potential need for ongoing model refinement.[13] The reliability of AI-generated content appears closely related to the credibility of referenced sources, as ChatGPT-4o and Gemini 2.0 predominantly relied on peer-reviewed journals or official guidelines. In contrast, Copilot and Perplexity often cited less reliable sources. An investigation focusing on epilepsy management observed that LLMs struggle to discern high-quality references, especially for conditions

Table 5 Relationship between response categories and EAACI compliance, reliability, agreement, and evidence levels.

	Median (IQR)	Mean Rank	WAO/EAACI-Compliant References	p*	WAO/EAACI Non-Compliant Reliable References	p*	Unreliable References	p*	Agreement Correlation	p**	Evidence Correlation	p**
First Responses												
ChatGPT	5 (4-5)	69.77	0 (0.0)	-	0 (0.0)	-	0 (0.0)	-	0.343762	0.073	-0.406739	0.032
Perplexity	5 (4-5)	70.32	22 (78.6)	0.920	28 (100.0)	-	14 (50.0)	0.112	0.339506	0.077	-0.291819	0.132
Gemini	5 (1.75-5)	54.27	3 (10.7)	0.900	2 (7.1)	0.248	0 (0.0)	-	0.502559	0.006	-0.001550	0.993
Copilot	3.0 (2-4)	31.64	22 (78.6)	0.565	19 (67.9)	0.136	13 (46.4)	0.204	0.200544	0.306	-0.317180	0.100
Second Responses												
ChatGPT	5 (4-5)	62.84	0 (0.0)	-	20 (71.4)	0.559	0 (0.0)	-	0.572871	0.001	-0.145880	0.459
Perplexity	5 (4-5)	65.59	18 (64.3)	0.841	28 (100.0)	-	15 (53.6)	0.881	0.615775	<0.001	-0.232615	0.234
Gemini	5 (5-5)	69.16	3 (10.7)	0.203	4 (14.3)	0.897	5 (17.9)	0.563	0.571649	0.001	-0.185529	0.345
Copilot	3 (2-4)	28.41	20 (71.4)	0.045	20 (71.4)	0.439	14 (50.0)	0.895	0.498282	0.007	-0.252668	0.195

*: Mann-Whitney U Test.
**: Spearman Correlation Analysis.

with scarce data.[10] This points to the broader necessity of refined citation prioritization systems that limit reliance on unverified sources.[16,17]

Another noteworthy aspect of this study is the role of contextual prompting. Gemini showed marked improvements when additional instructions were provided regarding reputable references, implying that more specialized guidance can enhance the accuracy and reliability of specific models. An analysis of the AI-driven breast cancer screening found that domain-specific training can lead to performance levels close to or even exceeding those of experienced clinicians, further supporting the concept that targeted fine-tuning is crucial for clinical reliability.[8] However, Copilot's consistent shortcomings indicate that not all models are equally responsive to refined prompts, and some may require reengineering or more robust training datasets to address the complexities of rare conditions like HAE adequately.

The findings also raise ethical and regulatory concerns. One article emphasized that transparency in AI decision-making is integral in preventing disparities in care, particularly in underserved populations at greater risk of algorithmic biases.[18] Unclear citation practices and limited insight into model architecture complicate the safe integration in routine workflows. Additionally, the evolving nature of LLMs introduces uncertainties over time, as periodic updates can modify performance in unpredictable ways. Without standardized benchmarks or independent regulatory oversight, clinicians may inadvertently rely on outdated or incomplete outputs, thereby compromising patient safety. This underscores the importance of ongoing validation efforts and caution in interpreting AI-derived recommendations.

There were several limitations in this study. The evaluation relied on 28 guideline-based questions, which may not fully capture the breadth of HAE clinical scenarios. The performance of LLMs is heavily influenced by the extent and quality of their training data, and rare diseases like HAE receive comparatively limited representation in these datasets.[19] Furthermore, the rapid pace of AI model updates makes it challenging to maintain consistent assessments over time. Additionally, guideline recommendations were reformulated into clear, structured questions for LLM evaluation. However, in actual clinical settings, questions posed by clinicians may often be less structured, ambiguous, or subject to misinterpretation, potentially limiting the generalizability of the current results to real-world clinical practice. The Likert-based scoring system offered a structured approach but may have overlooked subtle qualitative aspects of the models' recommendations. Despite these constraints, the study provides insights that could inform the refinement of AI tools and their integration into specialized clinical areas.

Looking toward clinically actionable AI, the results suggest that LLMs can significantly advance the dissemination of guideline-informed healthcare information, particularly for less common conditions where clinical expertise may be dispersed. However, genuine effectiveness can only be achieved if models undergo disease-specific fine-tuning and if developers establish transparent validation frameworks to enable real-time verification of recommendations. One investigation has advocated dedicated training programs to help clinicians critically evaluate AI-generated outputs, ensuring that clinical expertise remains the primary determinant of patient care.[20] Sustained regulatory oversight is also essential to address biases, safeguard data privacy, and manage unpredictable shifts in model capabilities that can occur with frequent updates. By incorporating these elements, LLMs could evolve into reliable adjuncts that complement rather than supplant the nuanced judgment of healthcare professionals in HAE management and beyond.

Conclusion

This study underscores the potential of LLMs in clinical practice, with ChatGPT and Gemini demonstrating stronger adherence to WAO/EAACI guidelines compared to Perplexity and Copilot. However, variability in citation reliability and reference quality highlights the need for continued refinement to ensure trustworthiness. While LLMs assures in disseminating guideline-aligned medical knowledge, their integration into healthcare requires targeted adaptations to address disease-specific complexities and rigorous validation protocols. Crucially, AI-generated recommendations must complement rather than replace clinician judgment to safeguard patient care. Future efforts should prioritize optimizing model accuracy, mitigating biases, and establishing standardized regulatory frameworks to expand their utility in rare disease management responsibly.

Author Contributions

All authors contributed equally.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to research, authorship, and/or publication of this article.

Funding

None.

References

1. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
2. Golan R, Reddy R, Ramasamy R. The rise of artificial intelligence-driven health communication. Transl Androl Urol. 2024;13:356–8. https://doi.org/10.21037/tau-23-556
3. Altıntaş E, Ozkent MS, Gül M, Batur AF, Kaynar M, Kılıç Ö, et al. Comparative analysis of artificial intelligence chatbot recommendations for urolithiasis management: A study of EAU guideline compliance. Fr J Urol. 2024;34(7–8):102666. https://doi.org/10.1016/j.fjurol.2024.102666
4. Reyhan AH, Mutaf Ç, Uzun İ, Yüksekyayla F. A performance evaluation of large language models in keratoconus:

A comparative study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. J Clin Med. 2024;13(21):6512. https://doi.org/10.3390/jcm13216512
5. Boyd CJ, Hemal K, Sorenson TJ, Patel PA, Bekisz JM, Choi M, et al. Artificial intelligence as a triage tool during the perioperative period: Pilot study of accuracy and accessibility for clinical application. Plast Reconstr Surg Glob Open. 2024;12(2):e5580. https://doi.org/10.1097/GOX.0000000000005580
6. Fu L, Kanani A, Lacuesta G, Waserman S, Betschel S. Canadian physician survey on the medical management of hereditary angioedema. Ann Allergy Asthma Immunol. 2018;121(5):598-603. https://doi.org/10.1016/j.anai.2018.06.017
7. Greve J, Lochbaum R, Trainotti S, Ebert EV, Buttgereit T, Scherer A, et al. The international HAE guideline under real-life conditions: From possibilities to limits in daily life - current real-world data of 8 German angioedema centers. Allergologie select. 2024;8:346-57. https://doi.org/10.5414/ALX02530E
8. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94. https://doi.org/10.1038/s41586-019-1799-6
9. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65.
10. Rocha-Silva R, de Lima BE, Costa TG, Morais NS, José G, Cordeiro DF, et al. Can people with epilepsy trust AI chatbots for information on physical exercise? Epilepsy Behav. 2024;163:110193. https://doi.org/10.1016/j.yebeh.2024.110193
11. Behers BJ, Stephenson-Moe CA, Gibons RM, Vargas IA, Wojtas CN, Rosario MA, et al. Assessing the quality of patient education materials on cardiac catheterization from artificial intelligence chatbots: An observational cross-sectional study. Cureus. 2024;16(9):e69996. https://doi.org/10.7759/cureus.69996
12. Maurer M, Magerl M, Betschel S, Aberer W, Ansotegui IJ, Aygören-Pürsün E, et al. The international WAO/EAACI guideline for the management of hereditary angioedema—the 2021 revision and update. Allergy. 2022;77(7):1961-90. https://doi.org/10.1111/all.15214
13. Tsai CY, Cheng PY, Deng JH, Jaw FS, Yii SC. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. Digit Health. 2024;10:20552076241269538. https://doi.org/10.1177/20552076241269538
14. Gokmen O, Gurbuz T, Devranoglu B, Karaman MI. Artificial intelligence and clinical guidance in male reproductive health: ChatGPT4.0's AUA/ASRM guideline compliance evaluation. Andrology. 2025;13(2):176-83. https://doi.org/10.1111/andr.13693
15. Birkun AA, Gautam A. Large language model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. Prehosp Disaster Med. 2023;38(6):757-63. https://doi.org/10.1017/S1049023X23006568
16. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: A systematic review and meta-analysis. Commun Med (Lond). 2024;4(1):71. https://doi.org/10.1038/s43856-024-00492-0
17. Olczak J, Pavlopoulos J, Prijs J, Ijpma FFA, Doornberg JN, Lundström C, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: An introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop. 2021;92(5):513-25. https://doi.org/10.1080/17453674.2021.1918389
18. Mirzaei T, Amini L, Esmaeilzadeh P. Clinician voices on ethics of LLM integration in healthcare: A thematic analysis of ethical concerns and implications. BMC Med Inform Decis Mak. 2024;24(1):250. https://doi.org/10.1186/s12911-024-02656-3
19. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: Development and validation. J Med Internet Res. 2023;25:e48763. https://doi.org/10.2196/48763
20. Deb Roy A, Bharat Jaiswal I, Nath Tiu D, Das D, Mondal S, Behera JK, et al. Assessing the utilization of large language model chatbots for educational purposes by medical teachers: A nationwide survey from India. Cureus. 2024;16(11):e73484. https://doi.org/10.7759/cureus.73484

Full Question List

1. Should all patients suspected of having HAE be assessed for blood levels of C1-INH function, C1-INH protein, and C4?
2. Should testing for C1-INH function, C1-INH protein, and C4 be repeated in patients who test positive to confirm the diagnosis of HAE-1/2?
3. Should patients suspected of having HAE and exhibiting normal C1-INH levels and function be assessed for known mutations underlying HAE-nC1-IN?
4. Should on-demand treatment be considered for all HAE attacks?
5. Should any HAE attack affecting or potentially affecting the upper airway be treated?
6. Is it recommended to treat HAE attacks as early as possible?
7. Should HAE attacks be treated with either intravenous C1 inhibitor, ecallantide, or icatibant?
8. Is it recommended to consider early intubation or surgical airway intervention in cases of progressive upper airway edema for HAE patients?
9. Is it necessary that all patients with HAE carry a sufficient number of on-demand medications at all times?
10. Is it recommended for HAE patients to consider short-term prophylaxis before medical, surgical, or dental procedures as well as exposure to other angioedema attack-inducing events?
11. Is the use of intravenous plasma-derived C1 inhibitor recommended as the first-line short-term prophylaxis for HAE patients?
12. Is it suggested for HAE patients to consider prophylaxis prior to exposure to patient-specific angioedema-inducing situations?
13. Are the goals of treatment recommended to be achieving total control of the disease and normalizing patients' lives for HAE patients?
14. Is it recommended for HAE patients to evaluate patients for long-term prophylaxis at every visit, considering disease activity, burden, and control as well as patient preference?
15. Is the use of plasma-derived C1 inhibitor recommended as the first-line long-term prophylaxis for HAE patients?
16. Is the use of lanadelumab recommended as the first-line long-term prophylaxis for HAE patients?
17. Is the use of berotralstat recommended as the first-line long-term prophylaxis for HAE patients?
18. Is the use of androgens recommended only as second-line long-term prophylaxis for HAE patients?
19. Is it suggested that all HAE patients using long-term prophylaxis be routinely monitored for disease activity, impact, and control to inform the optimization of treatment dosages and outcomes?
20. Is it recommended to carry out testing for children from HAE-affected families as soon as possible and to test all offspring of an affected parent?
21. Is it recommended to use C1 inhibitor or icatibant for the treatment of attacks in HAE diagnosed children under the age of 12?
22. Is it recommended to use plasma-derived C1 inhibitors as the preferred therapy during pregnancy and lactation for HAE patients?
23. Is it recommended for all HAE patients to have an action plan and treatment plan?
24. Is it recommended to have HAE-specific comprehensive, integrated care available for all HAE patients?
25. Is it recommended for HAE patients to be treated by a specialist with specific expertise in managing HAE?
26. Is it recommended that all HAE patients provided with on-demand treatment licensed for self-administration should be taught to self-administer?
27. Is it recommended that all HAE patients be educated about triggers that may induce attacks?
28. Is it recommended to screen family members of HAE patients for HAE?