Original Paper

Using a Large Language Model Chatbot for Allergen Avoidance Counseling in Contact Dermatitis: Quantitative Evaluation Study

Pawit Phadungsaksawasdi[1], MD, PhD; Thanida Kamolphasukchuen[1], MD; Poohglin Tresukosol[2], MD

[1]Division of Dermatology, Chulabhorn International College of Medicine, Thammasat University, Rangsit, Pathumthani, Thailand
[2]Allergic Contact Dermatitis Clinic and Occupational Medicine, Institute of Dermatology, Ratchathewi, Bangkok, Thailand

Corresponding Author:

Pawit Phadungsaksawasdi, MD, PhD
Division of Dermatology
Chulabhorn International College of Medicine, Thammasat University
99 Moo 18, Klongluang
Rangsit, Pathumthani 12120
Thailand
Phone: 662-5644440 ext. 4541
Email: pawit.p@tu.ac.th

Abstract

Background: AI is increasingly impacting health care. Large language models (LLMs), such as ChatGPT, have shown potential in supporting clinical decision-making. However, their application in dermatology, particularly in allergen avoidance counseling for contact dermatitis, remains underexplored.

Objective: This study aimed to demonstrate a potential use case for LLM responses in recommending allergen-free products for the management of contact allergies.

Methods: A total of 30 allergen groups included in the international standard patch test were assessed. ChatGPT (GPT-4) was used in this study from September 2023 to January 2024. For each allergen, ChatGPT was tasked with recommending 5 allergen-free products using 3 question formats of varying complexity. Responses were included if they specified a product and were manually classified by dermatologists into 5 categories: valid, invalid, cross-reaction, unknown, and unresponsive.

Results: ChatGPT provided 79.20% (1782/2250) valid responses, with 1.78% (40/2250) classified as invalid, 2% (45/2550) classified as cross-reactions, 8.67% (195/2250) classified as unknown, and 8.36% (188/2250) classified as unresponsive. The simplest question format yielded the highest proportion of valid recommendations (640/750, 85.33%). ChatGPT performed best when recommending cosmetics (1233/1365, 90.33%) and medications (198/240, 82.5%). However, lower validity rates were observed for recommendations involving certain sensitive allergens, including fragrance mixes and para-phenylenediamine (PPD), with frequent cross-reactive recommendations related to Peru balsam.

Conclusions: The version of ChatGPT evaluated in this study showed potential as a supplementary tool for allergen avoidance counseling. However, its recommendations require verification and should not replace clinician guidance or established allergen avoidance resources.

JMIR Dermatol 2026;9:e100364; doi: 10.2196/100364

Keywords: allergic contact dermatitis; allergen avoidance counseling; artificial intelligence; AI; ChatGPT; large language models

Introduction

Background

As the curtain rises on the AI era, AI technology is having an increasingly greater impact on our daily lives, especially in the field of medicine. The application of generative AI in dermatology has been increasingly investigated, including its use for screening, diagnostic assistance, and clinical evaluation [1]. Large language model (LLM)–based chatbots have shown promising accuracy across research fields [2] and, in health care, can assist experts by interpreting complex medical data, performing well on medical examinations, and responding accurately to a broad range of clinical questions across specialties [3-6].

Allergic contact dermatitis (ACD) is a common dermatologic disease caused by allergen exposure. Patch testing is the diagnostic gold standard, and allergen avoidance is the mainstay of treatment, making patient counseling essential [7]. However, counseling is often time-consuming and can overwhelm patients, thereby limiting their comprehension [8,9]. Difficult-to-read ingredient labels and complex allergen names further challenge both patients and clinicians when selecting appropriate products [10]. These barriers may reduce adherence to avoidance measures and worsen symptoms.

However, as previously mentioned, there are numerous limitations to allergen avoidance counseling for ACD. Additionally, although some recent research has indicated that LLM-based chatbots can improve patient education [11], it is still unclear how effectively ChatGPT performs in providing allergen avoidance strategies. The goal of this study was to explore the application of LLM-based chatbots in dermatology offices and evaluate their level of integration and accuracy. Due to its data-intensive and complex nature, contact dermatitis serves as an ideal model for assessing the usefulness of LLMs in dermatological practice. Previous research suggests that using specific question formats enhances the accuracy of LLMs [12-14], and improvements in models or training datasets could further enhance their performance and accuracy when responding to medical questions [15]. In this study, the goal was to evaluate, using 3 different detailed inquiries, the validity and possible limitations of LLM-generated responses when offering allergen-free product recommendations for patients with ACD. Therefore, LLMs could potentially serve as useful, time-saving, and affordable tools to assist with patient education in contact dermatitis clinics.

Objectives

This study aimed to demonstrate a possible use case for LLM-generated responses in recommending allergen-free products for the management of contact allergies and to assess the validity and limitations of this approach.

Methods

Data Collection and ChatGPT Inference

A total of 30 allergens from the Patch Test Series (IS1000, Chemotechnique Diagnostics) were included because this panel is widely used as a baseline for patch testing. For each allergen, 5 representative sources and 5 potentially allergen-containing products were identified from allergen information sheets. Between September 2023 and January 2024, all prompts were submitted manually through the ChatGPT web interface (OpenAI), using the model identified in the interface as GPT-4. Web browsing and retrieval were not enabled, and no external documents, plug-ins, or retrieval systems were provided. The exact model snapshot and system-level instructions were not visible through the interface, and inference parameters could not be specified.

Prompt Formats and Response Collection

For each allergen, GPT-4 was prompted to recommend 5 products considered suitable for patients with an allergy to that allergen (Figure 1). To assess response diversity and identify the most practical prompt, 3 prompt formats (Table 1) with varying levels of detail were evaluated [12-14]. Question 1 (Q1) was a basic and direct inquiry. Question 2 (Q2) was designed to simulate how a patient with contact dermatitis might ask a physician. Question 3 (Q3) included a clinician role and a patient-counseling context. The 3 formats differed concurrently in wording, length, level of detail, and clinical framing.

Prompts were entered manually, and GPT-4 memory was unavailable, so no information persisted across sessions. A new chat was started for each question to avoid contextual bias. Each of the 30 allergens was evaluated across 5 representative product types using 3 prompt formats, resulting in 450 prompts (30 allergens×5 product types×3 prompt formats). Each prompt requested 5 specific product recommendations, resulting in 2250 expected product-level outcomes. Each prompt format therefore comprised 150 prompts and 750 expected product-level outcomes. Each expected product recommendation was analyzed as a separate product-level outcome. Responses were included for ingredient verification only when ChatGPT named a specific product model or brand; otherwise, they were labeled “unresponsive.” Questions producing unresponsive outputs were re-entered once using the same prompt, but the effects of reprompting were not formally evaluated.

Figure 1. Flowchart of the study protocol. The study included 30 allergen groups from the international standard patch test. For each allergen, the 5 most common sources of exposure were selected. Three different question formats—simple, more detailed, and very specific—were used to encode prompts into ChatGPT. The chatbot was tasked with recommending 5 allergen-free products per query. Responses were then evaluated for validity by dermatologists and categorized into 5 groups: valid, invalid, unknown, unresponsive, and cross-reaction.

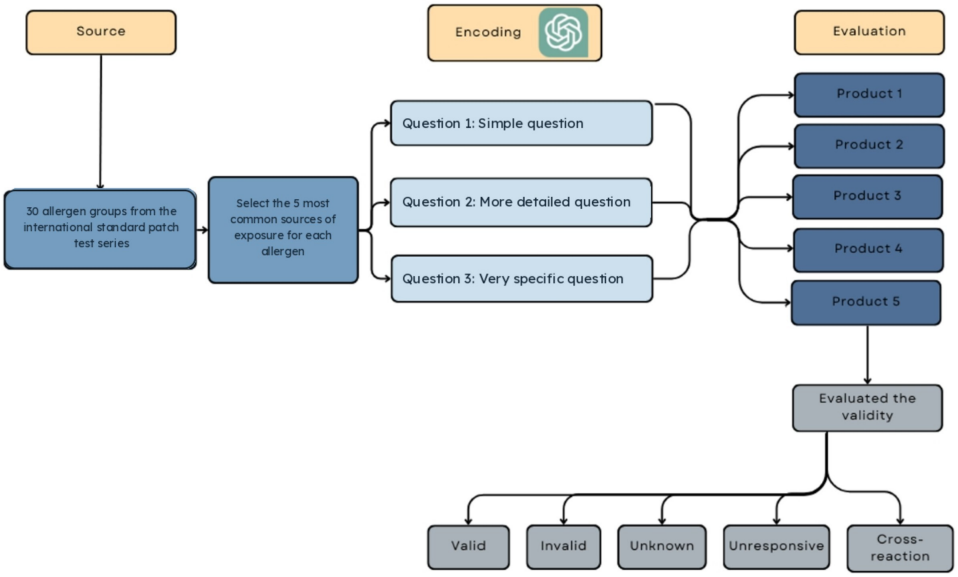


Table 1. Example of the prompt formats used in ChatGPT for allergen avoidance product recommendations[a].

Prompt formats	Prompts
Simple question	"Please recommend 5 specific (__Type of product__) brands that do not contain (__Name of allergen__)"
More detailed question	"If I have allergic to (__Name of allergen__), please recommend 5 specific (_Type of product_) brands that can be used in patients who are allergic to (__Name of allergen__)"
Very specific question	"You are the dermatologist, who works in the contact dermatitis clinic. There are many patients who test the patch test with you to investigate their contact dermatitis. After the patch test was done, you are the doctor who has to advise about the allergen's information and recommend a specific brand or product that is safe for your patients and does not contain the material that they are allergic to. Please recommend 5 specific (_Type of product_) that can be used by patients who are allergic to (__Name of allergen__)"

[a]This table presents 3 types of question prompts used to query ChatGPT for allergen-free product recommendations: a simple question, a more detailed question, and a very specific question. Each format was designed to reflect different levels of real-world clinical inquiry. The simple question is a direct request, the detailed question simulates a patient's inquiry, and the very specific question positions ChatGPT as a dermatologist providing professional counseling. Prompts are reproduced verbatim as entered during data collection; grammatical inconsistencies have therefore been retained.

Recommendation Classification and Verification

The validity of the answers was classified into 5 categories: "valid," "invalid," "unknown," "unresponsive," and "cross-reaction" (Figure 1). Products verified as not containing the specific allergen were labeled as valid, whereas products verified as containing the specific allergen were labeled as invalid. Products for which the ingredients could not be identified were labeled as "unknown," and products containing ingredients that cross-reacted with the allergen were labeled as "cross-reaction" (Multimedia Appendix 1).

The product-specific responses generated by ChatGPT were manually evaluated by 2 dermatologists. Each recommendation was reviewed by only 1 evaluator. For each set of 5 recommendations, 3 products were assigned to 1 dermatologist and 2 to the other, with this allocation alternated across questions. This approach was used to balance the overall assessment workload and ensure that both evaluators contributed to the evaluation of recommendations for each allergen. No individual recommendation was independently reviewed by both evaluators.

Product ingredients were verified using a predefined source hierarchy. The official manufacturer's website was considered the definitive source and was prioritized when information differed across sources. When manufacturer information was unavailable, safety data sheets were reviewed, followed by reliable third-party databases that compile product ingredient information: Environmental Working Group [16], What's in Products [17], FDA AccessData [18], SkinSAFE Products [19], INCIdecoder [20], and DailyMed [21].

Statistical Analysis

All responses were collected and analyzed in an Excel 2021 (version 2610; Microsoft Corp) spreadsheet. Descriptive results were reported as absolute numbers and percentages. To compare validity across the 3 prompt formats, the

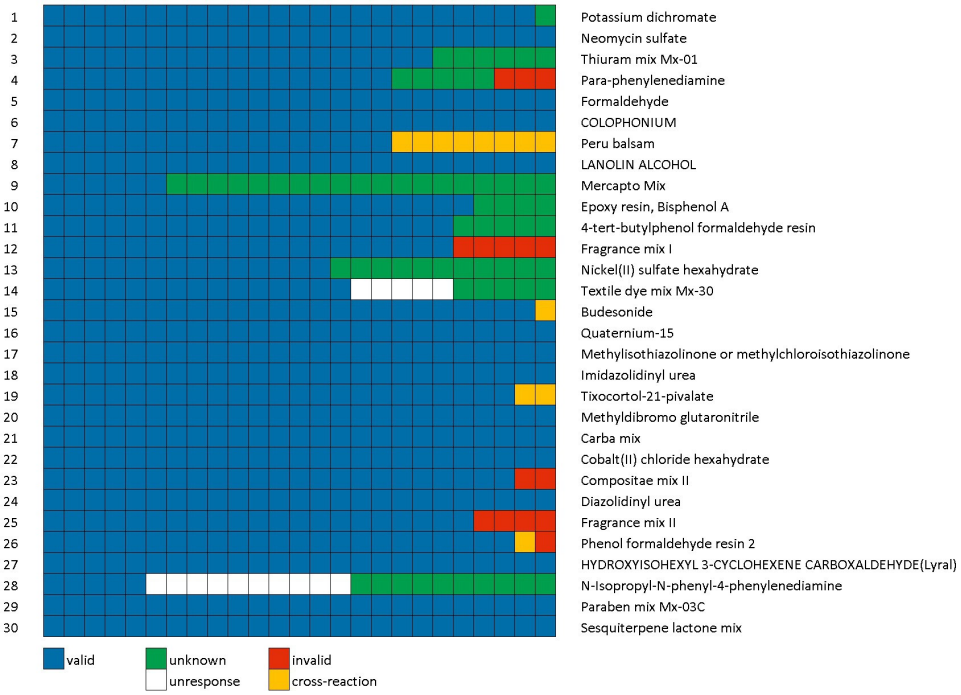
proportion of valid outcomes for each allergen was analyzed using the Friedman test, with the 30 allergens treated as paired units. When the overall test was significant, pairwise comparisons were performed using Wilcoxon signed-rank tests with Holm adjustment. A 2-sided *P* value of <.05 was considered statistically significant. Inferential analyses were performed using R (version 4.5.2; R Foundation for Statistical Computing).

Results

Overall Outcome Distribution

Across all 3 question formats, a total of 450 prompts were entered into ChatGPT on the first attempt. ChatGPT successfully provided product recommendations for 2062 (91.64%) out of 2250 products. Of these, 1782 (79.20%) responses were classified as valid (ie, the full ingredient lists were verified to be free of the specified allergens). A total of 40 (1.78%) responses were classified as invalid, as the recommended products contained the specific allergen or its components, and 45 (2%) responses involved ingredients known to cross-react with the specified allergens. In 195 (8.67%) cases, the ingredient list could not be identified. Additionally, 188 (8.36%) out of 2250 responses failed to provide a specific product brand or model. In these cases, ChatGPT gave general advice—such as suggesting that users check ingredient labels, contact manufacturers, or choose hypoallergenic or organic products—without naming specific product brands. The visualization of ChatGPT's responses is shown in Figure 2.

Figure 2. Visualization of the validity of ChatGPT responses to allergen-free product recommendations. This figure presents the results of ChatGPT's product recommendations for allergen avoidance across 30 allergens. Each horizontal row represents a specific allergen (listed on the right), and each box within the row corresponds to an individual response from ChatGPT. The color-coded boxes illustrate the validity of each recommendation: blue indicates a valid response, green indicates unknown ingredient information, red indicates an invalid response (containing the allergen), white indicates no specific product was recommended (unresponsive), and yellow indicates the presence of a cross-reactive allergen.



Thereafter, the questions that ChatGPT failed to answer during the first attempt were re-entered using the same input for a second attempt. The number of valid responses increased to 1871 (83.16%). However, the number of unknown responses and invalid responses also rose to 241 (10.71%) and 41 (1.82%), respectively. A total of 49 (2.18%) responses remained unresponsive.

Response Validity by Prompt Format

Among the 3 query formats, the simple question (Q1) yielded the highest percentage of valid responses, at 85.33% (640/750). The more detailed question (Q2) and the very specific question (Q3) resulted in 76.67% (575/750) and 75.60% (n=567) valid responses, respectively. In contrast to Q1, which had an unresponsiveness rate of only 2% (n=15), Q2 and Q3 had higher unresponsiveness rates of 12.13% (n=91) and 10.93% (n=82), respectively (Table 2). The overall difference in validity across prompt formats was statistically significant (Friedman test, *P*<.001), with Q1 showing higher validity than Q2 and Q3 (adjusted *P*=.01 for both comparisons), whereas Q2 and Q3 did not differ significantly (adjusted *P*=.49). Because the formats differed in several features, these observed differences cannot be attributed solely to prompt complexity or level of detail.

Table 2. ChatGPT's responses to 3 different question formats for each allergen from the international standard patch test (N=2250)[a].

Allergen names	Valid responses, % (n)			Total validity, % (n)	Invalid responses, % (n)			Unknown responses, % (n)			Unresponsive responses, % (n)			Cross-reactions, % (n)		
	Q1[b,c]	Q2[d,c]	Q3[e,c]	(Q1+Q2+Q3)[f]	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Potassium dichromate	96 (24)	96 (24)	92 (23)	94.67 (71)	0 (0)	0 (0)	0 (0)	4 (1)	4 (1)	8 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Neomycin sulfate	100 (25)	96 (24)	92 (23)	96 (72)	0 (0)	4 (1)	8 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Thiuram mix	76 (19)	72 (18)	64 (16)	70.67 (53)	0 (0)	0 (0)	0 (0)	24 (6)	28 (7)	36 (9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Para-phenylenediamine	68 (17)	68 (17)	64 (16)	66.67 (50)	12 (3)	12 (3)	12 (3)	20 (5)	20 (5)	24 (6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Formaldehyde	100 (25)	100 (25)	100 (25)	100 (75)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Colophonium	100 (25)	80 (20)	100 (25)	93.33 (70)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	0 (0)	0 (0)	0 (0)	0 (0)
Peru balsam	60 (15)	72 (18)	80 (20)	70.67 (53)	0 (0)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	40 (10)	24 (6)	20 (5)
Lanolin alcohol	100 (25)	100 (25)	100 (25)	100 (75)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0(0)	0 (0)
Mercapto mix	24 (6)	12 (3)	12 (3)	16 (12)	0 (0)	0 (0)	0 (0)	76 (19)	48 (12)	68 (17)	0 (0)	40 (10)	20 (5)	0 (0)	0 (0)	0 (0)
Epoxy resin, bisphenol A	84 (21)	80 (20)	48 (12)	70.7 (53)	0 (0)	0 (0)	0 (0)	16 (4)	16 (4)	24 (6)	0 (0)	4 (1)	28 (7)	0 (0)	0 (0)	0 (0)
4-tert-Butylphenol formaldehyde resin (para-tert-butylphenol)	80 (20)	76 (19)	36 (9)	64 (48)	0 (0)	0 (0)	0 (0)	20 (5)	24 (6)	4 (1)	0 (0)	0 (0)	60 (15)	0 (0)	0 (0)	0 (0)
Fragrance mix I	80 (20)	92 (23)	100 (25)	90.67 (68)	20 (5)	8 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Nickel(II) sulfate hexahydrate	56 (14)	48 (12)	44 (11)	49.33 (37)	0 (0)	0 (0)	0 (0)	44 (11)	32 (8)	36 (9)	0 (0)	20 (5)	20 (5)	0 (0)	0 (0)	0 (0)
Textile dye mix	60 (15)	60 (15)	56 (14)	58.67 (44)	0 (0)	0 (0)	0 (0)	20 (5)	12 (3)	44 (11)	20 (5)	28 (7)	0 (0)	0 (0)	0 (0)	0 (0)
Budesonide	96 (24)	96 (24)	72 (18)	88 (66)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	4 (1)	4 (1)	8 (2)
Quaternium-15	100 (25)	100 (25)	96 (24)	98.67 (74)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Methylisothiazolinone or methylchloroisothiazolinone	100 (25)	100 (25)	100 (25)	100 (75)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Imidazolidinyl urea	100 (25)	100 (25)	100 (25)	100 (75)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Tixocortol-21-pivalate	92 (23)	48 (12)	40 (10)	65.33 (45)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	40 (10)	40 (10)	8 (2)	12 (3)	20 (5)
Methydibromo glutaronitrile	100 (25)	80 (20)	100 (25)	93.33 (70)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	0 (0)	0 (0)	0 (0)	0 (0)
Carba mix	100 (25)	40 (10)	100 (25)	80 (60)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	60 (15)	0 (0)	0 (0)	0 (0)	0 (0)
Cobalt(II) chloride hexahydrate	100 (25)	100 (25)	80 (20)	93.33 (70)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	0 (0)	0 (0)	0 (0)
Compositae mix II	92 (23)	84 (21)	84 (21)	86.67 (65)	8 (2)	12 (3)	8 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (1)	8 (2)
Diazolidinyl urea	100 (25)	100 (25)	96 (24)	98.67 (74)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Fragrance mix II	84 (21)	84 (21)	92 (23)	86.67 (65)	16 (4)	16 (4)	8 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Phenol formaldehyde resin 2	92 (23)	40 (10)	76 (19)	69.33 (52)	4 (1)	4 (1)	0 (0)	0 (0)	4 (1)	0 (0)	0 (0)	52 (13)	20 (5)	4 (1)	0 (0)	4 (1)
Lyral	100 (25)	80 (20)	100 (25)	93.33 (70)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	0 (0)	0 (0)	0 (0)	0 (0)
N-Isopropyl-N-phenyl-4-phenylenediamine	20 (5)	12 (3)	0 (0)	10.67 (8)	0 (0)	0 (0)	0 (0)	40 (10)	28 (7)	40 (10)	40 (10)	60 (15)	60 (15)	0 (0)	0 (0)	0 (0)
Paraben mix	100 (25)	96 (24)	80 (20)	92 (69)	0 (0)	4 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	20 (5)	0 (0)	0 (0)	0 (0)

Allergen names	Valid responses, % (n)			Total validity, % (n)	Invalid responses, % (n)			Unknown responses, % (n)			Unresponsive responses, % (n)			Cross-reactions, % (n)		
	Q1[b,c]	Q2[d,c]	Q3[e,c]	(Q1+Q2+Q3)[f]	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Sesquiterpene lactone mix	100 (25)	88 (22)	64 (16)	84 (63)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	8 (2)	0 (0)	0 (0)	20 (5)	0 (0)	12 (3)	8 (2)

[a]This table presents the percentage of valid responses generated by ChatGPT for each allergen across 3 question formats: question 1 (simple), question 2 (more detailed), and question 3 (very specific). The "Total validity" column shows the average validity across the 3 question formats. Responses were also classified into 4 additional categories: invalid (recommendations containing the allergen), unknown (ingredient list could not be found), unresponsive (no specific product recommendations provided), and cross-reactions (recommendations containing known cross-reactive substances).

[b]Q1: question 1.

[c]Q1, Q2, and Q3 had 25 responses per allergen.

[d]Q2: question 2.

[e]Q3: question 3.

[f]Total validity was calculated from 75 responses per allergen.

Response Validity by Product Category

When evaluating ChatGPT's performance in providing product recommendations across various categories, particularly cosmetics, personal care products, and medicines, notable accuracy rates were observed. The results revealed that ChatGPT demonstrated a commendable validity in recommending relevant items within the cosmetics and personal care product categories, achieving an accuracy rate of 90.33% (1233/1365; Table 3). Moreover, in the medicine category, it also delivered recommendations with a high validity rate of 82.5% (198/240).

Table 3. Validity of ChatGPT's product recommendation responses by product category (N=2250)[a].

Response classification	Cosmetic and personal care products (n=1365), n (%)	Household products (n=300), n (%)	Construction and industrial materials (n=165), n (%)	Clothes and leather (n=135), n (%)	Pharmaceuticals (n=240), n (%)	Jewelry (n=45), n (%)
Valid	1233 (90.33)	214 (71.33)	130 (78.79)	0 (0)	198 (82.5)	7 (15.56)
Invalid	35 (2.56)	0 (0)	2 (1.21)	0 (0)	3 (1.25)	0 (0)
Cross-reaction	29 (2.12)	1 (0.33)	1 (0.61)	0 (0)	14 (5.83)	0 (0)
Unknown	28 (2.05)	48 (16)	3 (1.82)	88 (65.19)	0 (0)	28 (62.22)
Unresponsive	40 (2.93)	37 (12.33)	29 (17.58)	47 (34.81)	25 (10.42)	10 (22.22)

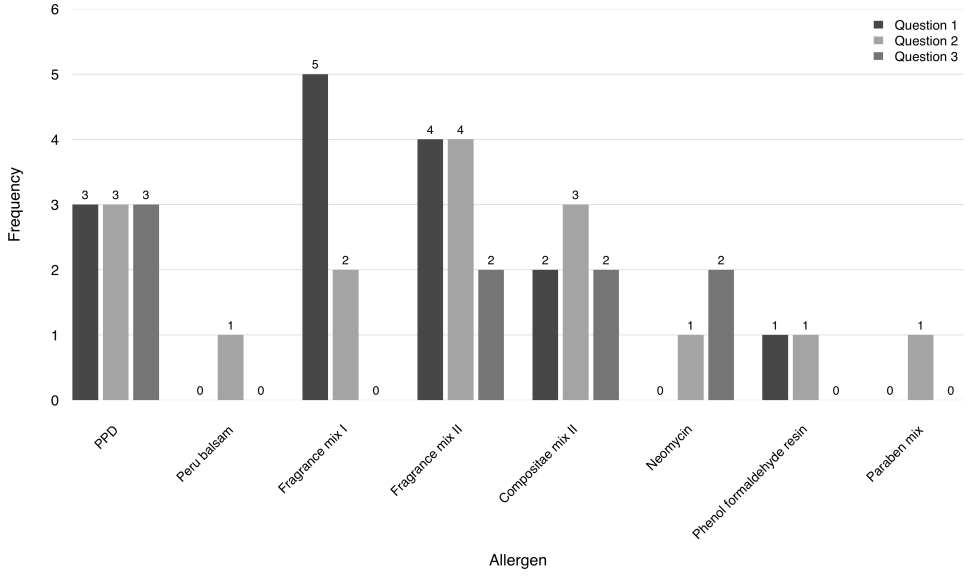
[a]This table summarizes the classification of product recommendations generated by ChatGPT across the 6 product categories. Responses were categorized as valid, invalid (containing the allergen), cross-reaction (containing cross-reactive substances), unknown (ingredient information could not be found), or unresponsive (no specific product was recommended).

For household products and construction materials, the model achieved validity rates of 71.33% (214/300) and 78.79% (130/165), respectively. However, for clothing and leather products, the model could not determine the validity of many recommendations, with 65.19% (88/135) of responses classified as unknown due to limitations in obtaining material details from manufacturers' websites. Obtaining ingredient information for products in the clothing and leather categories remained challenging, despite our best efforts to search for it. Accordingly, patients who are allergic to substances found in leather and textile materials may find it difficult to avoid exposure to these allergens.

Invalid Recommendations by Allergen

ChatGPT produced 40 invalid allergen avoidance recommendations, representing 1.78% (40/2250) of all responses. Fragrance mixes I and II and para-phenylenediamine (PPD), both common in cosmetics, accounted for most inaccuracies, with invalid recommendation rates of 11.33% (17/150) and 12% (9/75), respectively. Specifically, 9 (60%) of 15 hair dyes still contained PPD, and 1 of 15 perfumes included a Peru balsam component. Several recommended products also contained fragrance mix I—6 baby shampoos and 1 body lotion—or fragrance mix II—7 baby body washes, 2 antiperspirants, and 1 shampoo. In addition, 2 hair care products, 3 deodorants, and 2 moisturizers contained Compositae mix II, while 1 liquid soap contained paraben mix. In the medical category, inaccuracies included 3 medicines with neomycin, plus 1 lacquer and 1 anticorrosion product containing phenol formaldehyde resin (Figure 3).

Figure 3. Frequency of invalid product recommendations generated by ChatGPT by allergen and question format. This figure displays the number of invalid responses generated by ChatGPT for each allergen across the 3 question formats: simple (question 1), more detailed (question 2), and very specific (question 3). PPD: para-phenylenediamine.



Cross-Reactive Recommendations by Allergen

Among the 21 cosmetic and personal care products, including mouthwash, baby powder, shampoo, deodorant, and perfume, several were found to contain ingredients known to cross-react with Peru balsam, such as benzyl benzoate, benzyl salicylate, benzyl alcohol, eugenol, and benzoic acid. In addition, 14 medications were identified as containing ingredients with potential cross-reactivity to tixocortol-21-pivalate or budesonide. Cross-reactors associated with the Compositae mix and sesquiterpene lactone mix were also detected, albeit in a smaller proportion of cosmetic products.

Accordingly, the use of ChatGPT for allergen avoidance recommendations should be accompanied by caution regarding possible cross-reactions, particularly those involving Peru balsam and tixocortol-21-pivalate. It should also be noted that the definition of cross-reactivity may vary, and the criteria for including cross-reactors are detailed in the supplementary methods.

Discussion

Principal Findings

ACD management depends on identifying causative allergens and educating patients about how to avoid them. Patch testing and allergen avoidance counseling remain the cornerstones of treatment [22-24]. Effective counseling requires clear information on allergen sources, exposure routes, and alternative options [25]. However, this process is time-consuming, especially for patients sensitized to multiple allergens, and may be limited by access to trained clinicians or by patients' health literacy. AI has advanced rapidly in health care [26]. ChatGPT uses natural language processing to generate relevant responses [14]. Recent studies have demonstrated ChatGPT's promising performance in several medical domains, including cardiovascular prevention [4]. However, its applications in dermatology remain underexplored.

In this study, we provide the very first evidence that ChatGPT (GPT-4) can generate valid product recommendations for allergen avoidance in patients with ACD. According to our analysis, ChatGPT delivered valid responses in 79.20% (n=178) of cases, with 1.78% (n=4) classified as invalid and 8.36% (n=18) classified as unresponsive. When the same prompts were re-entered for a second attempt, the valid response rate increased to 83.16% (n=187), and the unresponsiveness decreased to 2.18% (n=4), indicating that repeated prompts may slightly enhance response accuracy. Of note, in practice, patients often present with multiple allergens. This study did not assess prompting with multiple allergens concurrently, which warrants further investigation. The variability in ChatGPT responses may be influenced by multiple factors, such as the timing and frequency of prompt entry. Previous studies by Ferreira et al [5] and Haze et al. [27] have shown that ChatGPT can produce varied outputs in response to identical prompts, especially when complex or multipart answers are required. Our findings align with these observations, although we observed that repeated prompts led to modest improvements in accuracy.

Response validity differed across the 3 tested prompt formats. The simple question produced the highest proportion of valid responses (640/750, 85.33%), compared with 76.67% (575/750) for the detailed question and 75.60% (567/750) for the very specific question. This suggests that the simple prompt format was associated with a higher proportion of valid responses than the 2 more detailed formats in the present study, consistent with the findings of Johnson et al [28]. However, the 3 formats varied simultaneously in wording, length, level of detail, and clinical framing. Therefore, these differences should not be interpreted as reflecting the isolated effect of prompt complexity or detail. ChatGPT performed particularly well in recommending allergen-free cosmetics, personal care items, and medicines, with validity rates of 90.33% (1233/1365) and 82.5% (198/240), respectively, for the combined cosmetics and personal care product categories and the medicine category, likely due to the availability of detailed ingredient information. In contrast, accuracy declined for household goods, construction materials, and especially clothing and leather products, where such information is often missing. Notably, inaccuracies were observed in recommendations involving certain allergens, such as fragrance mixes I and II and PPD. These results reflect only a subset of available product sources for each allergen.

Although all recommended products were searchable online and many were global brands, we did not assess local, in-market availability. This proof-of-concept study focused on the validity of LLM-provided allergen information rather than clinical applicability; however, real-world use would require regional adaptation to ensure that the recommended products are actually purchasable. Definitions of cross-reactors also vary among experts, and avoidance may follow standard or more restrictive approaches guided by physicians or allergen management programs. This variability influences the interpretation of results and highlights the need for greater standardization as such tools are developed.

The outcome categories have important implications for the clinical usefulness of the recommendations. Valid and invalid recommendations were defined by whether available ingredient information indicated the absence or presence, respectively, of the specified allergen. Valid recommendations therefore represent potentially suitable options, whereas invalid recommendations should not be considered suitable for patients with the specified allergy. Cross-reactive recommendations contained ingredients considered cross-reactive with the specified allergen and should not be regarded as allergen-free, highlighting the model's limited ability to account for clinically relevant cross-reactivity. Unknown recommendations could not be verified from available sources and should not be considered safe without further confirmation, such as updated manufacturer information or, when appropriate, chemical analysis. Although these outcomes may reflect limitations in available ingredient data, they also reduce the model's practical value for

allergen avoidance counseling. Unresponsive outputs did not provide a specific product recommendation and therefore offered limited actionable guidance. Overall, these categories highlight key gaps in ChatGPT-based counseling, including inadequate recognition of cross-reactivity, unverifiable ingredient information, and failure to provide specific product recommendations.

Importantly, general-purpose LLMs should not replace clinician counseling or dedicated allergen avoidance resources such as the Contact Allergen Management Program (CAMP). Although this study did not directly compare ChatGPT with CAMP, the invalid and cross-reactive recommendations observed indicate that the version of ChatGPT evaluated in this study should not be used as the sole source of information for selecting allergen-free products. Patients should verify LLM-generated recommendations using established allergen databases and current product labels or manufacturer information.

Limitations

This study has several limitations. First, we evaluated GPT-4 through the ChatGPT web interface during a defined period, and the findings may not generalize to other LLMs or future versions of the model. The exact model snapshot, system instructions, inference parameters, and backend updates were not available, limiting reproducibility. Given the nondeterministic and continuously evolving nature of LLMs, our results should be considered a time- and interface-specific assessment of ChatGPT performance. Future studies should use an API-accessible model with a documented snapshot and prespecified inference parameters. Because the prompt formats also differed in wording and grammatical structure, the observed differences in response validity may reflect these features in addition to differences in detail and clinical framing. Second, our scope was limited to a small set of allergens (30 allergens) and a limited range of product categories. Broader allergen and product coverage is needed, and regional differences in product availability may affect responses. Finally, technical settings such as temperature, which affects output variability, could not be adjusted in the web interface and should be addressed in future API-based work. Each product recommendation was assessed by only 1 of the 2 dermatologist evaluators; independent duplicate assessment was not performed. Consequently, evaluator disagreements could not be formally adjudicated, and interrater agreement could not be calculated. Although both evaluators applied the same classification criteria and source hierarchy, differences in product formulations and ingredient information across sources, regions, and time may have affected the classifications.

Conclusions

These findings suggest that LLM chatbots such as ChatGPT could support allergen avoidance management and help streamline contact dermatitis consultations. Because allergen management programs and databases may be restricted, centralized, and resource-intensive to maintain, LLM-based tools may offer an accessible supplementary source of information in settings with limited resources. Nevertheless, the version of ChatGPT evaluated in this study, as well as general-purpose LLMs not specifically validated for allergen avoidance, may recommend products containing the specified allergen or cross-reactive ingredients. Importantly, such tools should not replace standard allergen avoidance counseling but may assist clinicians and support decision-making. Clinicians should advise patients not to rely on LLM-generated recommendations alone and to confirm product suitability by consulting a dermatologist, using established resources such as the CAMP, and reviewing current product labels. This study may serve as a use-case demonstration to guide further development and evaluation of LLM-based tools in contact dermatitis.

Generative AI Use

During the preparation of this manuscript, the authors used ChatGPT (version 5.2; OpenAI) and Gemini (version 3; Alphabet Inc) to improve the language and style. These tools were not used to generate content, develop ideas, or write original sections of the text. Separately, ChatGPT was investigated as part of the study methodology and objectives, as detailed in the *Methods* section.

Funding

The authors declare that no funding was received for this study.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: PP
Data curation: TK, PP
Formal analysis: TK, PP
Methodology: PP
Supervision: PT, PP
Writing—original draft: TK, PP
Writing—review and editing: PP

Conflicts of Interest

None declared.

Multimedia Appendix 1

Representative case-level examples.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 1]

References

1. De A, Sarda A, Gupta S, Das S. Use of artificial intelligence in dermatology. Indian J Dermatol. 2020;65(5):352-357. [doi: 10.4103/ijd.IJD_418_20] [Medline: 33165383]
2. Azaria A, Azoulay R, Reches S. ChatGPT is a remarkable tool—for experts. Data Intell. Feb 1, 2024;6(1):240-296. [doi: 10.1162/dint_a_00235]
3. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. Mar 14, 2023;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 2023;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
5. Ferreira AL, Chu B, Grant-Kels JM, Ogunleye T, Lipoff JB. Evaluation of ChatGPT dermatology responses to common patient queries. JMIR Dermatol. Nov 17, 2023;6:e49280. [doi: 10.2196/49280] [Medline: 37976093]
6. Joly-Chevrier M, Nguyen AX, Lesko-Krleza M, Lefrançois P. Performance of ChatGPT on a practice dermatology board certification examination. J Cutan Med Surg. 2023;27(4):407-409. [doi: 10.1177/12034754231188437] [Medline: 37489920]
7. Mowad CM, Anderson B, Scheinman P, Pootongkam S, Nedorost S, Brod B. Allergic contact dermatitis: patient management and education. J Am Acad Dermatol. Jun 2016;74(6):1043-1054. [doi: 10.1016/j.jaad.2015.02.1144] [Medline: 27185422]
8. Jamil WN, Erikssohn I, Lindberg M. How well is the outcome of patch testing remembered by the patients? A 10-year follow-up of testing with the Swedish baseline series at the Department of Dermatology in Örebro, Sweden. Contact Dermatitis. Apr 2012;66(4):215-220. [doi: 10.1111/j.1600-0536.2011.02039.x] [Medline: 22404197]
9. Scalf LA, Genebriera J, Davis MD, Farmer SA, Yiannias JA. Patients' perceptions of the usefulness and outcome of patch testing. J Am Acad Dermatol. Jun 2007;56(6):928-932. [doi: 10.1016/j.jaad.2006.11.034] [Medline: 17239988]
10. Noiesen E, Munk MD, Larsen K, Johansen JD, Agner T. Difficulties in avoiding exposure to allergens in cosmetics. Contact Derm. Aug 2007;57(2):105-109. [doi: 10.1111/j.1600-0536.2007.01170.x]
11. Rames MM, O'Hern K, Rames JD, et al. ChatGPT as a resource for patient education in cosmetic dermatological procedures: a boon or a bane? J Cosmet Dermatol. Mar 2024;23(3):1085-1086. [doi: 10.1111/jocd.16166] [Medline: 38168902]
12. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. Dec 2023;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
13. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. J Eng Technol. 2023;43(1):52-68. [doi: 10.18260/jet.43.1.5]
14. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst. 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
15. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. Cardiovasc Intervent Radiol. Feb 2024;47(2):245-250. [doi: 10.1007/s00270-023-03563-2] [Medline: 37872295]
16. Environmental Working Group. URL: https://www.ewg.org/ [Accessed 2026-11-04]
17. What's in it? Consumer Product Information Database. URL: https://www.whatsinproducts.com/ [Accessed 2026-11-04]
18. Drug approvals and databases. U.S. Food & Drug Administration. URL: https://www.fda.gov/drugs/development-approval-process-drugs/drug-approvals-and-databases [Accessed 2026-11-12]
19. SkinSAFE. URL: https://www.skinsafeproducts.com/ [Accessed 2026-11-04]
20. INKEEDecoder. URL: https://inkeedecoder.com/ [Accessed 2026-11-04]
21. DailyMed. URL: https://dailymed.nlm.nih.gov/dailymed/ [Accessed 2026-11-04]
22. Belsito DV. The diagnostic evaluation, treatment, and prevention of allergic contact dermatitis in the new millennium. J Allergy Clin Immunol. Mar 2000;105(3):409-420. [doi: 10.1067/mai.2000.104937] [Medline: 10719287]
23. Johnston GA, Exton LS, Mohd Mustapa MF, et al. British Association of Dermatologists' guidelines for the management of contact dermatitis 2017. Br J Dermatol. Feb 2017;176(2):317-329. [doi: 10.1111/bjd.15239] [Medline: 28244094]

24. Steuer MS, Botto NC. Patient reported improvement after patch testing and allergen avoidance counseling: a retrospective analysis. Dermatol Ther (Heidelb). Sep 2018;8(3):435-440. [doi: 10.1007/s13555-018-0250-5] [Medline: 29987683]
25. Tramontana M, Hansel K, Bianchi L, Sensini C, Malatesta N, Stingeni L. Advancing the understanding of allergic contact dermatitis: from pathophysiology to novel therapeutic approaches. Front Med (Lausanne). 2023;10:1184289. [doi: 10.3389/fmed.2023.1184289] [Medline: 37283623]
26. Elechi U, Orobator ET, Udoh K, et al. Artificial intelligence in healthcare: a narrative review of recent clinical applications, implementation strategies, and challenges. J Healthc Leadersh. 2025;17:863-876. [doi: 10.2147/JHL.S553748] [Medline: 41488817]
27. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform. Dec 2023;180:105283. [doi: 10.1016/j.ijmedinf.2023.105283] [Medline: 37931432]
28. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. Feb 28, 2023:rs.3.rs-2566942. [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]

Abbreviations

ACD: allergic contact dermatitis
CAMP: Contact Allergen Management Program
LLM: large language model
PPD: para-phenylenediamine
Q1: question 1
Q2: question 2
Q3: question 3

Edited by Robert Dellavalle; peer-reviewed by John Meisenheimer, Kuan-Hsun Lin; submitted 05.May.2026; final revised version received 12.Oct.2026; accepted 13.Oct.2026; published 17.Nov.2026

Please cite as:
Phadungsaksawasdi P, Kamolphasukchuen T, Tresukosol P
Using a Large Language Model Chatbot for Allergen Avoidance Counseling in Contact Dermatitis: Quantitative Evaluation Study
JMIR Dermatol 2026;9:e100364
URL: https://derma.jmir.org/2026/1/e100364
doi: 10.2196/100364

© Pawit Phadungsaksawasdi, Thanida Kamolphasukchuen, Poohglin Tresukosol. Originally published in JMIR Dermatology (http://derma.jmir.org), 17.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Dermatology, is properly cited. The complete bibliographic information, a link to the original publication on http://derma.jmir.org, as well as this copyright and license information must be included.