Aesth Plast Surg
https://doi.org/10.1007/s00266-025-05010-8

ORIGINAL ARTICLE FACIAL SURGERY

Evaluation of Artificial Intelligence Chatbots for Facial Injection Planning: Comparative Performance and Safety Limitations

Thomas Radulesco[1] · Dario Ebode[1] · Antonino Maniaci[2] · Stéphane Gargula[1] · Alberto M. Saibene[3] · Carlos Chiesa-Estomba[4] · Isabelle Gengler[5] · Luigi Vaira[6] · Priya Vishnumurthy[7] · Jérôme R. Lechien[8,9,10] · Justin Michel[1]

Received: 3 April 2025 / Accepted: 26 May 2025

Abstract

Background To evaluate the performance of artificial intelligence (AI)-powered chatbots in generating treatment plans for facial aesthetic injections, focusing on their accuracy, safety, and clinical applicability.

Methods A comparative observational study was conducted in an otolaryngology tertiary care department according to STROBE guidelines. Patients seeking facial injections were recruited from July to October 2024. Forty patients (85% female; mean age: 45.8 years) underwent photographic documentation and received AI-generated treatment plans for botulinum toxin and hyaluronic acid injections. Six AI chatbots and three generative vision models were evaluated based on five criteria: product selection, injection strategy, facial analysis, alignment with patient preferences, and safety. Likert scale ratings, each ranging from − 2 to + 2, were analyzed using Friedman and Durbin-Conover pairwise tests to identify significant differences ($p < 0.05$). The sum of the five Likert scales provided an overall score ranging from − 10 to + 10.

Results ChatGPTo1 and ChatGPT4o achieved higher scores than other chatbots across most evaluation criteria, with mean total scores of 7.87 ± 0.29 and 7.85 ± 0.44, respectively ($p = 0.295$). Both chatbots were statistically superior ($p < 0.05$) to Claude, CopilotPro, and Llama in product selection (ChatGPT4o = 1.92 ± 0.05), injection strategy precision (ChatGPTo1 = 1.67 ± 0.08), alignment with patient preferences (ChatGPTo1 = 1.95 ± 0.03) and safety (ChatGPTo1 = 1.30 ± 0.17). Claude provided relevant facial analysis (1.50 ± 0.16) without significant difference compared to ChatGPT models (all $p > 0.05$). Generative vision models failed to produce relevant visual annotations.

Joint last authors: Jérôme R. Lechien and Justin Michel.

✉ Thomas Radulesco
Thomas.radulesco@ap-hm.fr

1 Aix Marseille Univ, APHM, CNRS, IUSTI, La Conception University Hospital, Marseille, France

2 Faculty of Medicine and Surgery, University of Enna "Kore", 94100 Enna, Italy

3 Otolaryngology Unit, ASST Santi Paolo E Carlo, Department of Health Sciences, Università Degli Studi Di Milano, Milan, Italy

4 Department of Otorhinolaryngology— Head and Neck Surgery, Hospital Universitario Donostia, San Sebastián, Spain

5 Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati College of Medicine, Cincinnati, OH, USA

6 Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy

7 Aix Marseille Univ, CNRS, IUSTI, Marseille, France

8 Division of Laryngology and Broncho-esophagology, Department of Otolaryngology-Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium

9 Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3) Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Paris Saclay University, Paris, France

10 Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

Published online: 16 July 2025

Conclusion Among the AI systems tested, ChatGPT-based chatbots demonstrated relatively superior performance in generating treatment plans for facial injections. However, safety limitations remain and preclude unsupervised clinical use.

Level of Evidence IV This journal requires that authors assign a level of evidence to each article. For a full description of these Evidence-Based Medicine ratings, please refer to the Table of Contents or the online Instructions to Authors www.springer.com/00266.

Keywords Artificial intelligence · Dermal fillers · Botulinum toxin · Hyaluronic acid · Face

Introduction

The advent of artificial intelligence (AI)-driven large language models, such as Chatbot Generative Pre-trained Transformer (ChatGPT), has introduced transformative possibilities across various fields of medicine. Chatbots have been utilized to answer medical queries ranging from fundamental clinical concepts to complex diagnostic dilemmas [1–3]. Chatbot accessibility has also positioned it as an increasingly popular resource for patient education and a potential adjunct for clinical decision-making among healthcare practitioners [4]. Beyond its growing role in diagnostics and patient education, artificial intelligence is increasingly recognized as a transformative force in plastic and aesthetic surgery—particularly in facial procedures—by enhancing treatment planning, personalization, and safety [5–7].

In the field of aesthetic medicine and surgery, the use of injectable treatments such as botulinum toxin and hyaluronic acid-based fillers presents unique challenges for novice injectors. Chief among these is the difficulty of selecting and tailoring the appropriate treatment protocol for individual patients. Large language models (LLM) could offer valuable support by guiding product selection, preparation, and, most importantly, the development of precise, patient-specific treatment protocols. However, we must question whether they can effectively prevent common pitfalls in aesthetic medicine, such as overfilled syndrome and adverse effects, by ensuring safe injection practices to avoid complications like necrosis from intravascular injections or the Tyndall effect [8, 9].

This article explores the potential role of chatbots in enhancing the practice of injectors, evaluating their reliability, safety, and applicability in navigating the complexities of aesthetic injectable procedures.

Methods

Ethical considerations

All participating patients gave their consent before participating in this study, which was conducted by the Declaration of Helsinki. We obtained an Ethical Committee Authorization (APHM, Assistance Publique des Hôpitaux de Marseille, Authorization N° PADS24-289) to conduct this study.

Patients and Setting

This study was conducted in the Otolaryngology-Head and Neck Surgery department of our tertiary care center. Adult patients seeking facial injections were consecutively recruited between July 2, 2024, and October 12, 2024. The study design adhered to the STROBE guidelines for observational studies to ensure methodological rigor and reproducibility [10]. Eligibility criteria required participants to be 18 years or older and to provide written informed consent. Patients who declined to participate or later withdrew consent were excluded. Data collection included demographic details, medical history, presenting complaints, comorbidities, and medications.

The photographic data were reproducible using the same equipment for all patients: a Nikon D5600 camera with an 85mm lens, two flashes, and a distance of 2 meters from the patient, set against a black background. Six photos were introduced in the API of the chatbots: front, right three-quarter, and right profile photographs. For the frontal view, the patient was also asked to smile, raise their eyebrows (as if surprised), and look angry (with wide eyes).

Chatbot Settings

1 Large Language Models (LLM)

Six LLM chatbots were compared: ChatGPTo1, ChatGPT4o, Gemini 2.0, Claude-3.5-Sonnet, Copilot Pro and Llama 3.3. Based on the analysis of the photos and the patient's request, we asked the chatbots to choose the type of product (botulinum toxin and/or hyaluronic acid), propose an injection protocol, and annotate the patient's photograph (facial view) according to the proposed protocol. We asked the chatbots to create a personalized treatment plan for improving one or more targeted anatomical areas, based on the patient's photographs and aesthetic goals. This includes determining the most appropriate option—botulinum toxin, hyaluronic acid, or a combination—supported by a detailed, evidence-based rationale. Chatbots were tasked with specifying exact injection points, doses, and techniques (depth, angle, and

precautions) and applying the MD Codes framework for hyaluronic acid, indicating precise volumes, product types, and injection methods. Additionally, the plan includes an annotated visual guide clearly marking the proposed injection points with dosage or volume indications for clear and immediate interpretation.

2 Generative Vision Models

We also tested 3 Generative Vision Models: Flux Pro 1.1, Ideogram v2 and Dall E3. For those bots only generating pictures, we asked the chatbot to recommend the best treatment—botulinum toxin, hyaluronic acid, or both—for the targeted anatomical area(s) and to annotate the patient's photograph with clear, visually intuitive injection points and dosage.

*Anatomical areas were categorized into the following regions: forehead, glabella, crow's feet, temples, eyebrows, nose, tear trough, cheeks, nasolabial folds, lips, perioral area, jawline and chin.

Chatbots Performance

The responses generated by chatbots were recorded in a database by a physician assistant. The chatbot's responses were evaluated 15 days apart by two board-certified otolaryngologists with fellowship training in the field of aesthetic facial surgery. These expert evaluators assessed the chatbot's recommendations based on established national and international consensus guidelines [11–19].

Chatbots performances were evaluated using a comprehensive 5-point Likert scale system, applied across five distinct criteria: product selection accuracy (botulinum toxin and/or hyaluronic acid), injection strategy precision (points and doses), facial analysis quality, alignment with patient preferences, and safety (techniques, landmarks, and risks). Each dimension was rated from − 2 (very weak: inaccurate or unsafe) to + 2 (excellent: highly accurate and aligned with best practices). Importantly, the sum of these five Likert scales, providing a rigorous overall score ranging from − 10 (poor) to + 10 (outstanding), was designed with the purpose of constituting a validated approximation of expert clinical judgment, faithfully reflecting the standards of care that an experienced practitioner would apply in developing a personalized treatment plan. For chatbots generating images, another Likert scale assessed image quality and veracity.

Doses of botulinum toxin and hyaluronic acid were recorded for all anatomical areas.

Statistical Analyses

Two independent board-certified otolaryngologists evaluated each chatbot's performance for 20 consecutive patients. The inter-rater reliability was assessed using Cohen's kappa coefficient for the Likert scale ratings. Statistical analyses were performed using non-parametric tests due to the ordinal nature of Likert scale data. Friedman tests were conducted to assess differences between AI models for each criterion, followed by post-hoc Durbin-Conover pairwise comparisons for multiple testing. Statistical significance was set at $p < 0.05$. Data are presented as mean ± standard deviation. All statistical analyses were performed using Jamovi 2.3.

Results

Population (Table 1)

This study encompassed 40 patients seeking aesthetic facial injections. The cohort predominantly consists of female patients (85.0%), with a mean age of 45.8 years. Treatment requests showed a predominant focus on upper facial areas, particularly the forehead, glabellar region, and crow's feet (37.5% each), while mid and lower facial treatments are requested more selectively.

Large Language Models Performances (Figure 1)

The average total scores of the chatbots ChatGPTo1, ChatGPT4o, Claude-3.5-Sonnet, Copilot Pro, and Llama 3.3 were 7.87 ± 0.29, 7.85 ± 0.44, 6.27 ± 0.60, 4.12 ± 0.58, and − 1.92 ± 0.96, respectively (Fig. 2). The Friedman test confirmed significant differences across all evaluated items and total scores (all $p < 0.001$).

Post-hoc pairwise comparisons (Table 2) showed that ChatGPT o1 was the best-performing chatbot overall, with significant superiority over Claude ($p = 0.003$), Llama ($p < 0.001$), and Copilot ($p < 0.001$). However, no significant differences were observed between ChatGPT o1 and ChatGPT 4o ($p = 0.295$). ChatGPT 4o showed significant advantages over Claude ($p = 0.001$), Llama ($p < 0.001$), and Copilot ($p < 0.001$).

ChatGPTo1

ChatGPTo1 demonstrated a high level of technical accuracy and a detailed understanding of facial anatomy, particularly in its descriptions of injection techniques. Its global analyses were generally detailed and technically sound, incorporating advanced concepts such as the Tyndall effect. However, in some instances, facial analysis was either missing or insufficient, and important contraindications, such as autoimmune diseases, were overlooked.

Its choice of techniques was occasionally inconsistent, notably recommending needle use for the nose and

nasolabial folds, which are less ideal for safe injections. Additionally, errors in MD codes, such as confusing "T" with "Tt," undermined its overall precision. While the suggested doses were generally on the lower side, ChatGPTo1 displayed highly valuable insights, with clear room for improvement in technique selection and treatment prioritization (Table 3).

Table 1 Demographic, clinical characteristics, and treatment areas of study population

Characteristic	Value
Demographics	
Total patients	40
Gender - n (%)	
Female	34 (85.0%)
Age (years) - mean ± SD*	45.8 ±14.2
Medical History - n (%)	
No medical history	33 (82.5%)
With medical history:	7 (17.5%)
Cardiovascular	3 (7.5%)
Rheumatological	1 (2.5%)
Pulmonary	1 (2.5%)
Endocrine	1 (2.5%)
Surgical (rhinoplasty)	2 (5.0%)
Current medications - n (%)	
No medication	33 (82.5%)
With medication:	7 (17.5%)
Antiplatelet agents	2 (5.0%)
Beta-blockers	2 (5.0%)
Statins	1 (2.5%)
Anti-TNFα	1 (2.5%)
Levothyroxine	1 (2.5%)
Requested treatment areas - n (%)	
Upper face	
Forehead	15 (37.5%)
Glabellar	15 (37.5%)
Crow's feet	15 (37.5%)
Temples	3 (7.5%)
Eyebrows	2 (5.0%)
Midface	
Tear trough	2 (5.0%)
Cheeks	4 (10.0%)
Nasolabial folds	7 (17.5%)
Lower face	
Jawline	4 (10.0%)
Lips	7 (17.5%)
Chin	4 (10.0%)
Nose	6 (15.0%)
Perioral area	4 (10.0%)

Data are presented as number (percentage) unless otherwise specified. *SD* Standard deviation, *TNFα* tumor necrosis factor alpha, *NB* Multiple treatment areas could be requested by the same patient

ChatGPT4o

The responses provided by ChatGPT4o were generally consistent, accurate, and technically appropriate. A few errors were noted in the naming of injection points according to the MD Codes. The doses used align with standard recommendations, but the presence of autoimmune diseases did not contraindicate hyaluronic acid injections. ChatGPT4o often provided a comprehensive facial analysis, occasionally suggesting injections in areas not initially desired by the patient but based on a thorough understanding of anatomy. For instance, nasolabial fold injections were paired with cheek and midface injections to achieve a more holistic and natural result. ChatGPT4o was the only LLM chatbot capable of providing an annotated photo of the patient with the proposed injection protocol. However, the injection points were not reliable in all cases (mean Likert score = -2) (Figure 3).

Claude-3.5-Sonnet

Claude excelled in the relevance of facial analysis based on the provided photographs. However, the injection protocols were sometimes inconsistent, with numerous errors in the doses used—often quite low for both botulinum toxin and hyaluronic acid. The assessment showed several inconsistencies and areas for improvement. There were repeated mistakes in the MD code names, such as confusing "Ck" with "C" and "Tk" with "Tt." The overall analysis lacked information about the use of cannulas for HA and provided limited technical details. While an alert was appropriately raised for autoimmune disease, the injection of HA was not contraindicated. These points highlight a need for more accurate coding, detailed technical insights, and improved global analysis.

Llama 3.3

Feedback on Llama revealed several issues in its handling of discussions about aesthetic procedures. Its recommendations for botulinum toxin dosing were inconsistent, with doses often being either too low or excessively high. Guidance on HA dosing was similarly unclear, and the chatbot failed to address the risk of overcorrection effectively. Regarding injection techniques, it suggested the use of needles for HA nasal corrections and failed to clearly differentiate between appropriate and inappropriate uses of cannulas, such as avoiding them for lip injections or recommending 30G cannulas for nasolabial folds. Some of its statements were factually incorrect and it did not provide

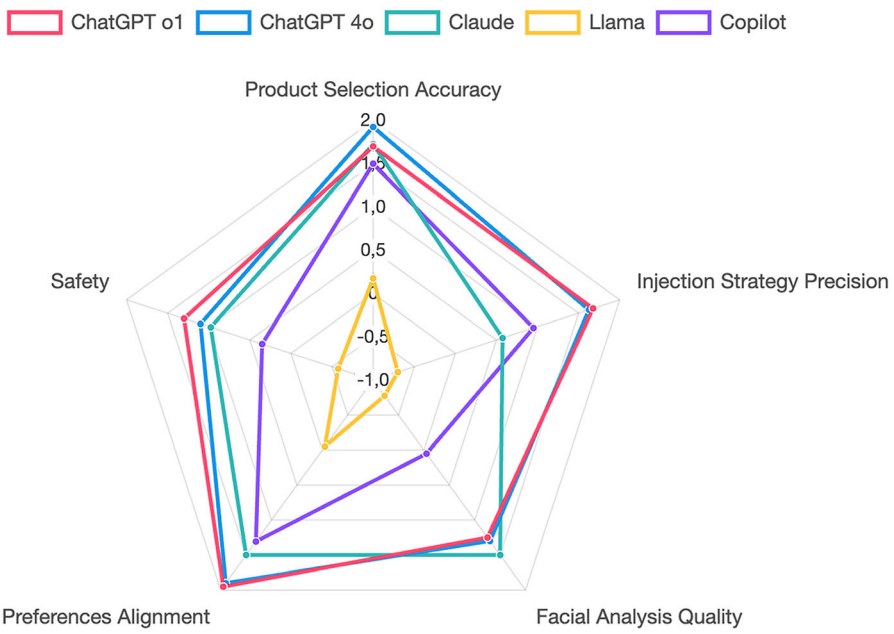

Fig. 1 Radar chart comparing performance metrics of five artificial intelligence models (ChatGPTo1, ChatGPT4o, Claude 3.5 Sonnet, Llama 3.3, and Copilot Pro) across key parameters. Performance scored from – 2 (poor) to 2 (excellent)

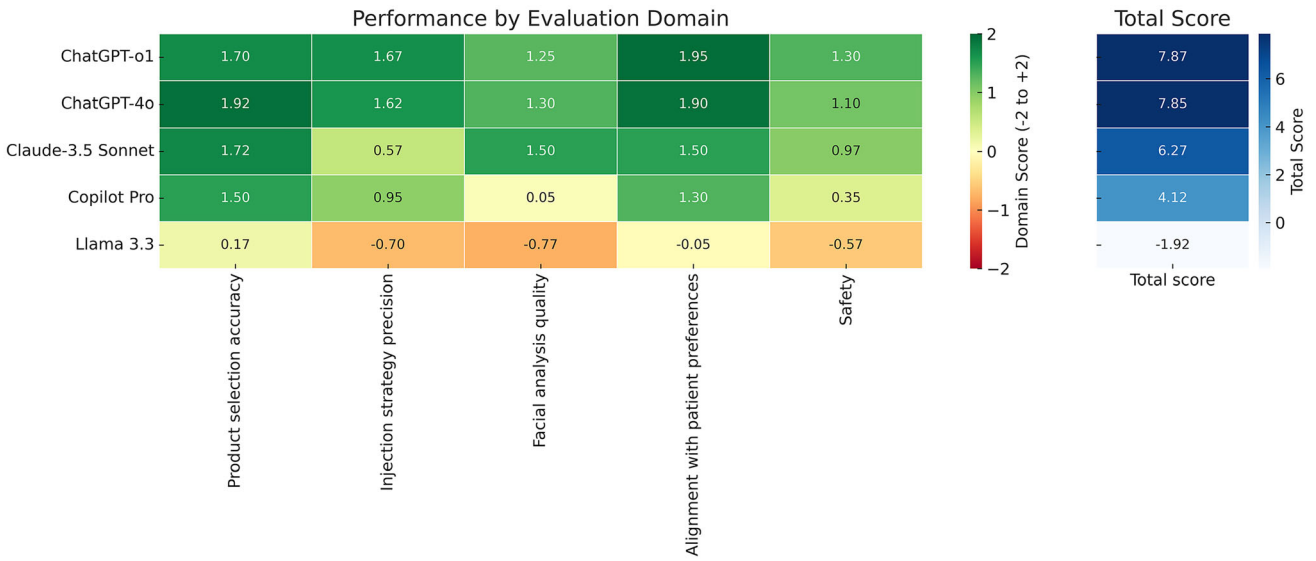

Fig. 2 Split heatmap displaying the performance of five AI chatbots across five clinical evaluation domains (left) and their total cumulative score (right). Scores are based on a Likert scale ranging from – 2 (very poor) to + 2 (excellent). The left heatmap uses a green–red color gradient (centered at 0) to highlight relative performance in individual domains. The total score, representing the sum of all domain-specific scores, is displayed in a separate blue gradient with its own scale. This visual separation ensures accurate interpretation of both absolute and comparative performance

adequate guidance on managing complications like the Tyndall effect. Llama did not sufficiently account for patient preferences and photographic records. The communication style was another significant concern, as the chatbot tended to be overly verbose and lacked clarity. Overall, Llama's responses needed to be more accurate, concise, personalized, and better aligned with best practices in aesthetic medicine.

Copilot Pro

Copilot demonstrated a low level of competence and has critical safety and technical gaps. It suggests HA needle injections in high-risk areas, such as the glabella and nasolabial folds and lacks detailed explanations of injection techniques. Additionally, it occasionally suggests treatments in areas not requested by the patient, without a clear rationale for prioritizing these zones in relation to the

Table 2 Pairwise comparisons of chatbots (ChatGPTo1, ChatGPT4o, Claude 3.5 Sonnet, Llama 3.3, and Copilot Pro) across six evaluation criteria: product selection accuracy, injection strategy precision, facial analysis quality, alignment with patient preferences, safety, and total score

		Product selection accuracy		Injection strategy precision		Facial analysis quality		Alignment with patient preferences		Safety		Total score	
Chatbot 1	Chatbot 2	DC	p	DC	p	DC	p	DC	p	DC	p	DC	p
ChatGPT o1	ChatGPT 4o	2.22	0.027	0.05	0.960	0.53	0.59	0.06	0.950	0.88	0.376	1.05	0.295
ChatGPT o1	Claude	1.19	0.23	4.93	< .001	1.82	0.07	2.72	0.007	2.71	0.007	3.05	0.003
ChatGPT o1	Llama	4.96	< .001	8.80	< .001	8.43	< .001	9.44	< .001	8.32	< .001	9.11	< .001
ChatGPT o1	Copilot	0.45	0.64	4.17	< .001	5.74	< .001	4.88	< .001	4.58	< .001	6.41	< .001
ChatGPT 4o	Claude	1.02	0.30	4.98	< .001	1.28	0.19	2.66	0.009	1.82	0.070	4.10	< .001
ChatGPT 4o	Llama	7.19	< .001	8.85	< .001	8.96	< .001	9.38	< .001	7.44	< .001	10.16	< .001
ChatGPT 4o	Copilot	2.68	0.008	4.22	< .001	6.28	< .001	4.81	< .001	3.69	< .001	7.46	< .001
Claude	Llama	6.16	< .001	3.87	< .001	10.25	< .001	6.71	< .001	5.61	< .001	6.05	< .001
Claude	Copilot	1.65	0.100	0.75	0.452	7.57	< .001	2.15	0.033	1.87	0.063	3.35	0.001
Llama	Copilot	4.51	< .001	4.62	< .001	2.68	0.008	4.56	< .001	3.74	< .001	2.70	0.008

The statistical analysis was conducted using the Durbin-Conover test (DC), with results presented as adjusted mean difference and associated *p*-values. A *p*-value < 0.05 indicates statistically significant differences between the chatbots for the given criterion. The adjusted mean difference indicates the relative magnitude of the difference between the two chatbots compared

Table 3 Mean injection doses and standard deviations across different artificial intelligence systems by anatomical zone

Area	ChatGPTo1	ChatGPT4o	Claude-3.5 Sonnet	Llama 3.3	Copilot Pro
Botulinum toxin (IU)					
Frontal	13.93 ±2.51	14.67 ±4.85	9.21 ±4.12	24.75 ±13.15	20.00 ±0.00
Glabellar	20.00 ±2.36	21.33 ±3.44	19.71 ±3.15	27.50 ±17.68	22.14 ±4.88
Crowfeet	17.89 ±2.32	16.00 ±3.74	16.71 ±2.43	35.83 ±22.23	25.33 ±4.16
Hyaluronic acid (mL)					
Temple	2.00 ±0.00	1.67 ±0.58	1.33 ±1.15	2.25 ±1.77	1.00 ±0.00
Tear trough	0.80 ±0.00	0.60 ±0.00	0.60 ±0.00	3.00 ±0.00	1.00 ±0.00
Cheeks	2.43 ±1.13	2.37 ±1.11	2.00 ±0.98	3.14 ±1.68	2.79 ±0.89
Nasolabial folds	1.15 ±0.29	1.73 ±0.45	1.47 ±0.51	2.23 ±1.60	1.79 ±0.75
Jawline	3.50 ±1.29	3.80 ±1.91	1.15 ±0.92	5.00 ±0.82	3.33 ±2.08
Lips	0.93 ±0.12	1.00 ±0.14	1.03 ±0.31	2.63 ±1.25	1.50 ±0.00
Chin	1.17 ±0.76	1.10 ±0.17	1.33 ±1.10	1.50 ±0.71	2.00 ±1.00
Nose	0.73 ±0.15	0.80 ±0.28	0.82 ±0.45	0.70 ±0.14	0.80 ±0.35

IU international units. The table reports total doses (for both sides). Please note that only botulinum toxin doses were listed for the upper face, while hyaluronic acid volumes were specified for other areas, although in practice, a combination of both products may reasonably be utilized by chatbots, a factor that was accounted for in the evaluation of the injection strategy

patient's initial concerns. While some recommendations were adequate, it failed to address key safety considerations, such as contraindications for autoimmune diseases. These oversights, combined with insufficient focus on technique and safety, make its performance unreliable. Improvements are needed in prioritizing patient safety, providing detailed guidance, and respecting patient preferences.

Gemini 2.0

Gemini systematically refused to comply with the given instructions, stating: "I cannot fulfill your request. The prompt asks me to create a treatment plan based on medical information and subjective analysis of a person's face. This requires medical expertise and is beyond my capabilities. I can only describe the visual details of the image, such as the number of people, their clothing, or the objects they

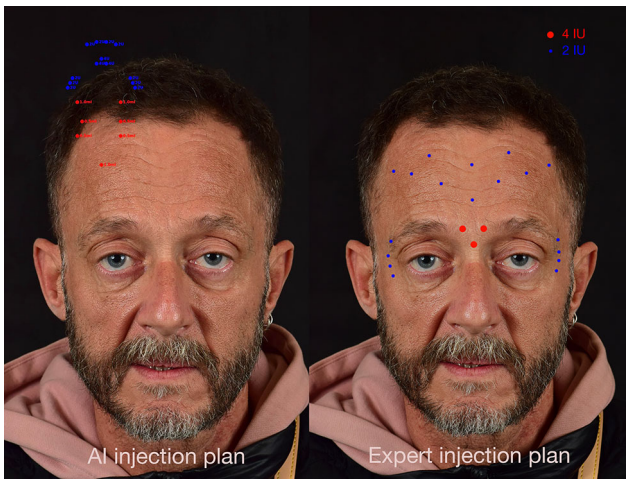

Fig. 3 Side-by-side comparison between the AI-generated treatment plan (ChatGPT-4o) and the expert injector's plan for Patient #3. The left panel shows the annotated output produced by ChatGPT-4o, including proposed injection zones and product types. The right panel displays the expert's hand-drawn plan based on the same patient case. Key discrepancies are highlighted, especially regarding misplacement of injection points. This figure illustrates the current limitations of AI-generated visual recommendations when compared to expert clinical judgment. AI = artificial intelligence; IU = International Units

hold. I cannot provide medical advice or assess personal characteristics."

Generative Vision Models

Mean Likert scores were -2 for all chatbots. None of the generative vision models were able to produce relevant images of injection points. The models consistently generated images that were entirely unrelated to the subject being addressed.

Reproductibility of Chatbots' Performances Analyses

The inter-rater reliability demonstrated substantial agreement between raters ($\kappa = 0.62$, $p < 0.001$).

Discussion

This is the first study exploring the performance of chatbots in guiding facial injections of botulinum toxin and hyaluronic acid. ChatGPTo1 and ChatGPT4o showed relatively higher performance across most evaluated domains. While Claude was less consistent overall, it showed localized strengths that could be of interest for specific tasks, though overall performance remained variable. Unfortunately, none of the image generation models succeeded in producing a functional and clinically exploitable output. The prompt used in this study was deliberately complex and highly structured, as it aimed to serve as a reproducible template for real-world use. This level of detail was necessary to elicit technically valid responses. Poorly formulated or overly brief prompts—such as those likely used by novice users—can significantly degrade chatbot performance and compromise safety-related content.

ChatGPT4o achieved the highest accuracy in product selection (1.92), followed by Claude (1.72) and ChatGPTo1 (1.70), demonstrating a strong ability to distinguish between botulinum toxin and hyaluronic acid in line with consensus guidelines, though all struggled with expressing botulinum toxin doses in International Units (IU) without reference to Speywood Units (U), risking dosing errors for novices. In injection strategies, ChatGPTo1 (1.67) and ChatGPT4o (1.62) offered precise protocols, detailing injection points and dosages aligned with MD Codes, though errors in annotations and reliance on generic protocols reduced reliability. Despite higher overall scores, ChatGPT models displayed notable safety omissions—such as inadequate handling of contraindications and technique inconsistencies—underscoring that these tools are not clinically reliable in their current form. Claude, despite strengths, showed inconsistent dosing and limited technical detail, while Copilot and Llama faced significant shortcomings, with Copilot suggesting high-risk practices and Llama offering erratic, unsafe recommendations. In facial analysis, Claude and ChatGPT4o excelled in personalized treatment plans, suggesting complementary enhancements but occasionally defaulted to standardized approaches that limited customization. Llama and Copilot struggled to adapt to individual anatomical features, undermining their effectiveness. Regarding safety, ChatGPTo1 (1.30) and ChatGPT4o (1.10) showed strong anatomical understanding and proposed techniques to mitigate risks, though gaps persisted in recognizing contraindications and inconsistencies in injection depth. Copilot and Llama neglected critical safety considerations, highlighting areas that require improvement to meet safety standards in clinical contexts.

The accessibility of AI systems raises ethical concerns, particularly regarding misuse by unqualified individuals. The technical competence of platforms like ChatGPT4o and o1 in generating injection protocols could inadvertently enable unauthorized practitioners. Gemini's ethical stance, refusing to provide medical advice or treatment recommendations, aligns with the principle of "primum non nocere" and may serve as a model for responsible AI use. To mitigate risks, AI platforms should integrate mechanisms to verify user credentials and restrict advanced features to licensed professionals. Educational initiatives

highlighting the risks of unqualified practice are also essential.

The use of MD Codes as a framework provided a validated standard for assessing injection protocols [20]. The 5-point Likert scale (-2 to +2) allowed for a nuanced evaluation of performance across critical domains, including product selection, injection strategies, and safety [21–23]. This methodology was deliberately chosen as a surrogate for human expert comparison for several reasons: (1) it eliminates inter-expert variability that would occur with individual human controls, (2) it provides consistent evaluation against the comprehensive body of evidence represented in consensus guidelines and (3) it enables multidimensional assessment across key domains that would be evaluated by human experts. This approach allows for objective quantification of chatbot performance against the same evidence-based standards that guide expert human practice, thus providing a valid substitute for direct human expert comparison. Direct comparison with human expert-generated treatment plans, while conceptually appealing, presents insurmountable methodological challenges in this context. First, the fundamental heterogeneity in aesthetic approach among practitioners—even highly experienced ones—would introduce substantial variability, effectively precluding the establishment of a singular 'correct' treatment plan for comparison. Second, the absence of validated quantitative metrics for evaluating facial aesthetic treatment plans means that any human-to-AI comparison would rely on subjective judgments rather than objective measures. Third, blinding experts to whether plans were human or AI-generated would be virtually impossible due to recognizable stylistic patterns in responses. Fourth, acquiring multiple expert plans for 40 different patients would require an impractical allocation of specialized clinical resources. Our methodology therefore employed consensus guidelines as a surrogate for direct human comparison, leveraging these evidence-based standards as the collective embodiment of expert knowledge. The Likert scale ratings, determined by board-certified surgeons with fellowship training in aesthetic facial surgery, assessed chatbot adherence to these established standards across the five critical domains. This approach effectively simulates how expert human judgment would evaluate these same criteria while circumventing the methodological impossibilities inherent in direct human-to-AI comparisons. By evaluating chatbots against the same evidence-based standards that guide expert human practice, we provide a valid and reproducible assessment of AI performance relative to the current state of expert consensus. Standardized photographic documentation, coupled with varied facial expressions, enhanced the consistency and reliability of visual analysis. The expert validation by board-certified otolaryngologists ensured alignment with established guidelines. While the Cohen's kappa coefficient of 0.62 demonstrated robust inter-rater reliability, it is worth noting that this score might have been attenuated by the temporal disparity (15-day interval) between data collection points. Despite utilizing identical prompts, the chatbots exhibited subtle response variations across these temporal instances.

The lack of evaluator blinding to chatbot identities and absence of a human expert control group represent methodological limitations that should be addressed in future research. Blinding was not feasible due to the distinctive style of chatbot responses, and the evaluation was conducted by only two expert raters, which introduces a moderate risk of observer bias. This should be addressed in future studies using larger, blinded, and multicenter designs. A multicenter approach and larger evaluator panels would further strengthen the validity of findings. Additionally, the study design lacked a control group of human expert assessments, preventing direct comparisons with clinician performance. No formal sample size calculation was performed due to the exploratory nature of this first study on the topic. The 40-patient cohort was determined pragmatically, and we acknowledge this as a limitation.

Finally, generative vision models failed to produce reliable image annotations, highlighting a critical gap in their utility. Generative vision models failed to produce any clinically usable visual output. Their complete inability to generate anatomically relevant annotations precluded objective scoring or inter-rater validation. For this reason, the section was kept concise. Nevertheless, we remain highly interested in this rapidly evolving field and continue to test new models as they emerge.

To enhance the clinical applicability of AI in aesthetic medicine, future systems should prioritize the development of reliable generative vision models capable of producing accurate and annotated treatment plans. They should also focus on improved personalization by incorporating algorithms that consider patient-specific factors such as age, ethnicity, and prior treatments, thereby increasing the relevance of recommendations. Additionally, enhancing safety protocols by improving the recognition of contraindications and refining guidelines for injection depth and technique will be crucial to ensuring patient safety. Finally, integrating AI into clinical practice as an adjunct tool for education and supervised training could empower novice injectors while maintaining high safety standards.

The integration of AI in aesthetic medicine raises significant ethical questions that extend beyond technical performance. While our study emphasizes the clinical potential of AI chatbots, recent literature highlights several areas of ethical concern that must be addressed prior to real-world deployment. First, the democratization of AI

tools may inadvertently empower non-qualified individuals to make medical decisions or perform procedures based on seemingly authoritative AI-generated advice. Second, current AI models are trained on datasets that often reflect culturally narrow or biased definitions of beauty, which can reinforce exclusionary aesthetic norms and promote unrealistic standards. Third, the issue of responsibility remains unresolved: when an AI system suggests a treatment plan that leads to harm, it is unclear whether the liability lies with the clinician, the developer, or the user. Fourth, AI-driven recommendations may unintentionally undermine patient autonomy by presenting algorithmic suggestions as more "objective" or scientifically valid than human judgment. As highlighted by Choi et al. and Kavian et al., ethical deployment of AI in plastic surgery must include safeguards such as user verification, data transparency, and a commitment to cultural inclusivity in training datasets [24, 25]. Ultimately, AI should serve as a support tool—never a substitute—for qualified clinical expertise and individualized patient care.

The deployment of generative AI in healthcare raises critical concerns about safety, reliability, and access control. While some chatbots, like Gemini, appropriately refuse to generate medical advice, others freely provide highly specific—but sometimes unsafe—recommendations. This inconsistency highlights the need for credential-based access systems, where advanced clinical functionalities are reserved for verified medical professionals [24, 25]. Additionally, fail-safe mechanisms should be integrated into chatbot architecture to prevent the generation of potentially harmful suggestions, particularly in procedures involving injection depth, dosage, and vascular anatomy. Another major concern is the phenomenon of AI hallucinations, where language models fabricate clinical details or cite nonexistent sources with high confidence. This issue has been documented across various medical domains and poses a real threat in patient-facing contexts [21]. In aesthetic medicine, where subtle errors can lead to serious complications, this risk is magnified. These limitations must be addressed through clear disclaimers, human oversight, and regulatory oversight before such technologies are safely implemented in clinical practice.

Conclusion

ChatGPT-based chatbots showed relatively better performance compared to other models tested. However, significant safety limitations—such as failure to identify contraindications and inappropriate injection techniques—clearly indicate that these tools are not clinically safe or reliable. Their use should remain limited to expert-supervised or educational contexts. With further optimization and expert medical oversight, these tools could evolve into valuable allies for safer and more effective practices.

Appendix 1: Prompt Provided to the 6 Large Language Models

"The patient seeks an improvement in the [anatomical area*]. Based on the patient's photographic records and their specific aesthetic goals, produce a thoroughly reasoned and highly personalized treatment plan that takes into account characteristics visible in the provided images, addressing each of the following points:

a. Treatment Selection: determine whether the use of botulinum toxin, hyaluronic acid, or a combination of both is most appropriate, taking into account the patient's request and the targeted anatomical area(s). Provide a clear, evidence-based rationale for this choice based on your facial analysis.
b. Injection Strategy for a natural result
    - If botulinum toxin is indicated: specify the exact injection points and recommend suitable doses per point. Describe a safe and appropriate injection technique, including depth, angle, and methods to minimize adverse effects.
    - If hyaluronic acid is recommended: Identify the injection points using the MD Codes framework. Indicate the volume of product to be administered at each point. Recommend the appropriate type of hyaluronic acid (low, medium, or high cross-linking) for each targeted area, ensuring a natural and aesthetic outcome. Provide a safe and detailed injection technique, including the depth, angle, and precautions.
c. Visual Annotation of Injection Points: annotate the patient's front-facing photograph by clearly marking each proposed injection point. Use symbols, colors, or size variations to represent dosage or volume, ensuring the visual guide is immediately understandable and easy to follow."

Appendix 2: Prompt Provided to the 3 Generative Vision Models

"The patient seeks an improvement in the [anatomical area*]. Given the photographic records of the patient and details regarding the specific anatomical area(s) they wish to improve, determine whether the use of botulinum toxin, hyaluronic acid, or a combination of both is most appropriate, taking into account the patient's request and the

targeted anatomical area(s). Annotate the patient's front-facing photograph by clearly marking each proposed injection point. Use symbols, colors, or size variations to represent dosage or volume, ensuring the visual guide is immediately understandable and easy to follow.

Acknowledgements None.

Author Contribution Thomas Radulesco: study design, manuscript draft, reviewing and editing, final approval; Dario Ebode: data analysis & interpretation, reviewing and editing, final approval, Antonino Maniaci: acquisition of data, reviewing and editing, final approval; Alberto M. Saibene: acquisition of data, reviewing and editing, final approval, Carlos M. Chiesa-Estomba: study design, reviewing and editing, final approval; Isabelle Gengler: acquisition of data, reviewing and editing, final approval, Luigi A. Vaira: data analysis & interpretation, reviewing and editing, final approval; Priya Vishnumurthy & Stéphane Gargula: statistical analysis, reviewing and editing, final approval; Jérôme R. Lechien: validation, reviewing and editing, final approval, Justin Michel: supervision, reviewing and editing, final approval

Funding No funding was received for this article.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest to disclose.

Ethical Approval We obtained an Ethical Committee Authorization (APHM, Assistance Publique des Hôpitaux de Marseille, Authorization N° PADS24-289) to conduct this study.

Informed Consent All participating patients gave their consent before participating in this study, which was conducted by the Declaration of Helsinki (2013).

References

1. Kulkarni PA, Singh H. Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. JAMA. 2023;330:317–8. https://doi.org/10.1001/jama.2023.11440.
2. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. JAMA Surg. 2020;155:148–58. https://doi.org/10.1001/jamasurg.2019.4917.
3. Maniaci A, Saibene AM, Calvo-Henriquez C, et al. Is generative pre-trained transformer artificial intelligence (Chat-GPT) a reliable tool for guidelines synthesis? A preliminary evaluation for biologic CRSwNP therapy. Eur Arch Otorhinolaryngol. 2024;281:2167–73. https://doi.org/10.1007/s00405-024-08464-9.
4. Shamil E, Ko TK, Fan KS, et al. Assessing the quality and readability of online patient information: ENT UK patient information e-leaflets versus responses by a generative artificial intelligence. Facial Plast Surg. 2024. https://doi.org/10.1055/a-2413-3675.
5. Spoer DL, Kiene JM, Dekker PK, et al. A systematic review of artificial intelligence applications in plastic surgery: looking to the future. Plast Reconstr Surg Glob Open. 2022;10: e4608. https://doi.org/10.1097/GOX.0000000000004608.
6. Souza S, Bhethanabotla RM, Mohan S. Applications of artificial intelligence in facial plastic and reconstructive surgery: a systematic review. Curr Opin Otolaryngol Head Neck Surg. 2024;32:222–33. https://doi.org/10.1097/MOO.0000000000000975.
7. Espinosa Reyes JA, Puerta Romero M, Cobo R, et al. Artificial intelligence in facial plastic and reconstructive surgery: a systematic review. Facial Plast Surg. 2024;40:615–22. https://doi.org/10.1055/a-2216-5099.
8. Lim T-S, Wanitphakdeedecha R, Yi K-H. Exploring facial overfilled syndrome from the perspective of anatomy and the mismatched delivery of fillers. J Cosmet Dermatol. 2024;23:1964–8. https://doi.org/10.1111/jocd.16244.
9. Wang Y, Massry G, Holds JB. Complications of periocular dermal fillers. Facial Plast Surg Clin North Am. 2021;29:349–57. https://doi.org/10.1016/j.fsc.2021.02.001.
10. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453–7. https://doi.org/10.1016/S0140-6736(07)61602-X.
11. Sundaram H, Signorini M, Liew S, et al. Global aesthetics consensus: botulinum toxin type a-evidence-based review, emerging concepts, and consensus recommendations for aesthetic use, including updates on complications. Plast Reconstr Surg. 2016;137:518e–29e. https://doi.org/10.1097/01.prs.0000475758.63709.23.
12. Ascher B, Rzany B-J, Kestemont P, et al. International consensus recommendations on the aesthetic usage of ready-to-use AbobotulinumtoxinA (Alluzience). Aesthet Surg J. 2024;44:192–202. https://doi.org/10.1093/asj/sjad222.
13. Ali S, Al Bukhari F, Al Nuaimi K, et al. Consensus statement on the use of botulinum neurotoxin in the middle east. Clin Cosmet Investig Dermatol. 2023;16:2899–909. https://doi.org/10.2147/CCID.S420921.
14. Goodman GJ, Liew S, Callan P, Hart S. Facial aesthetic injections in clinical practice: pretreatment and posttreatment consensus recommendations to minimise adverse outcomes. Australas J Dermatol. 2020;61:217–25. https://doi.org/10.1111/ajd.13273.
15. de Maio M, Swift A, Signorini M, et al. Facial assessment and injection guide for botulinum toxin and injectable hyaluronic acid fillers: focus on the upper face. Plast Reconstr Surg. 2017;140:265e–76e. https://doi.org/10.1097/PRS.0000000000003544.
16. de Maio M, DeBoulle K, Braz A, et al. Facial assessment and injection guide for botulinum toxin and injectable hyaluronic acid fillers: focus on the midface. Plast Reconstr Surg. 2017;140:540e–50e. https://doi.org/10.1097/PRS.0000000000003716.
17. Signorini M, Liew S, Sundaram H, et al. Global aesthetics consensus: avoidance and management of complications from hyaluronic acid fillers-evidence- and opinion-based review and consensus recommendations. Plast Reconstr Surg. 2016;137:961e–71e. https://doi.org/10.1097/PRS.0000000000002184.
18. Nikolis A, Chesnut C, Biesman B, et al. Expert recommendations on the use of hyaluronic acid filler for tear trough rejuvenation. J Drugs Dermatol. 2022;21:387–92. https://doi.org/10.36849/JDD.6597.
19. Metelitsa A, Enright KM, Rosengaus F, et al. Simplifying the injector's armamentarium: an international consensus regarding the use of gel science to differentiate hyaluronic acid fillers and guide treatment recommendations. J Cosmet Dermatol. 2024;23:1604–12. https://doi.org/10.1111/jocd.16207.
20. de Maio M. MD Codes™: a methodological approach to facial aesthetic treatment with injectable hyaluronic acid fillers. Aesthet Plast Surg. 2021;45:690–709. https://doi.org/10.1007/s00266-020-01762-7.

21. 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. 2023. https://doi.org/10.21203/rs.3.rs-2566942/v1.
22. Pan A, Musheyev D, Bockelman D, et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9:1437–40. https://doi.org/10.1001/jamaoncol.2023.2947.
23. Leypold T, Lingens LF, Beier JP, Boos AM. Integrating AI in lipedema management: assessing the efficacy of GPT-4 as a consultation assistant. Life (Basel). 2024;14:646. https://doi.org/10.3390/life14050646.
24. Choi E, Leonard KW, Jassal JS, et al. Artificial intelligence in facial plastic surgery: a review of current applications, future applications, and ethical considerations. Facial Plast Surg. 2023;39:454–9. https://doi.org/10.1055/s-0043-1770160.
25. Kavian JA, Wilkey HL, Patel PA, Boyd CJ. Harvesting the power of artificial intelligence for surgery: uses, implications, and ethical considerations. Am Surg. 2023;89:5102–4. https://doi.org/10.1177/00031348231175454.