

Article

Repeatability of Artificial Intelligence Chatbots in Composite Shade Selection: Agreement with a Dental Specialist

Seyit Bilal Ozdemir ^{1,*} , Busra Ozdemir ¹  and Cagri Ural ² ¹ Department of Restorative Dentistry, Faculty of Dentistry, Giresun University, Giresun 28200, Turkey² Department of Prosthodontics, Faculty of Dentistry, Giresun University, Giresun 28200, Turkey

* Correspondence: seyit.ozdemir@giresun.edu.tr; Tel.: +90-454-310-1690; Fax: +90-454-310-1691

Abstract

This study aimed to evaluate the intra-model repeatability of three artificial intelligence-based chatbots (ChatGPT-4.0, Microsoft Copilot, and Claude 3.5) in composite shade selection and their agreement with a dental specialist. Ten acrylic resin maxillary central incisor teeth representing different VITA Classical shades ($n = 10$) were photographed together with A1, A2, and A3 composite shade tabs under standardized illumination. Shade selections were performed by each artificial intelligence model based on the photographs and repeated on five different days using identical images and prompts. Visual shade selection by the dental specialist was determined by consensus between two calibrated evaluators. CIE L^* , a^* , and b^* values of the acrylic teeth and composite shade tabs were obtained by photometric analysis, and color differences were calculated using the CIEDE2000 formula. Intra-model repeatability was assessed using Fleiss' kappa coefficient, and agreement with the dental specialist was evaluated using Cohen's kappa statistic. Intra-model repeatability differed among the models, with ChatGPT-4.0 demonstrating fair repeatability ($\kappa = 0.33$), Claude 3.5 showing moderate repeatability ($\kappa = 0.45$), and Microsoft Copilot exhibiting poor repeatability ($\kappa = -0.12$). Trial-level agreement with the dental specialist varied across repeated assessments, with ChatGPT-4.0 generally demonstrating higher agreement than the other models, whereas Microsoft Copilot showed consistently low agreement. Artificial intelligence chatbots showed variable repeatability and limited agreement with expert evaluation in composite shade selection under standardized conditions.

Keywords: artificial intelligence; composite resin; dental photography; esthetic

1. Introduction

In esthetic restorative dental treatments, selecting a shade that harmonizes with the natural tooth structure is a critical factor for both clinical success and patient satisfaction [1]. Shade guides are commonly used for composite restorations; however, their use may lead to errors in color matching, primarily because these guides are typically made of ceramic materials that possess different optical properties compared to composite resins [2,3]. An alternative approach involves applying composite buttons directly onto relevant areas of the intact tooth surface and polymerizing them, which can serve as a reliable method for shade selection [2,4]. Nevertheless, this technique is limited by unpredictable effects such as layering and color changes due to the hydration of the composite material [5]. The development of custom-made shade guides with various composite color combinations and thicknesses may support this process [2,6]. Shade selection largely relies on the clinician's visual judgment, which is inherently subjective and influenced by several external factors,



Academic Editor: Kathrin Becker

Received: 8 February 2026

Revised: 24 February 2026

Accepted: 25 February 2026

Published: 27 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

including ambient lighting, operator experience, visual fatigue, and individual perception differences [7–9].

Artificial intelligence is the imitation of human mental processes by machines [10]. In recent years, advancements in artificial intelligence applications, particularly in natural language processing and image analysis, have enabled these systems to contribute not only to text-based recommendations but also to clinical decision-making processes [11]. Recent studies have examined the performance of Artificial Intelligence Chatbots in various medical fields such as emergency medicine, pathology, radiology, surgical education, and dietetics [12–15]. In the field of dentistry, a growing number of studies have demonstrated the potential effectiveness of artificial intelligence in diagnosis, disease prevention, and treatment planning [16,17]. Additionally, the performance of Artificial Intelligence Chatbots in evaluating knowledge within dental education has been investigated [18,19]. At present, artificial intelligence technologies are utilized in various diagnostic procedures, such as identifying carious lesions in radiographs [20], evaluating the difficulty level of endodontic treatments [21] and categorizing different dental implant systems [22]. However, overlooking the limitations of artificial intelligence in critical fields such as medicine and dentistry may lead to serious concerns [23]. The accuracy and reliability of smartphone-assisted tooth shade selection were assessed in a recent *in vitro* study [24]. Another study has evaluated the reliability, consistency, and readability of responses provided by Artificial Intelligence Chatbots to questions related to restorative dentistry [25]. Recent dentistry-focused studies have begun to examine the reliability of Large Language Model-based chatbots in clinically sensitive contexts, particularly in medication-related decision support and the identification of drug–drug interactions [26,27]. These investigations underscore the importance of evaluating the consistency and reliability of artificial intelligence systems when they are applied to dental decision-making processes. In a recent study evaluating the accuracy of color matching techniques using Artificial Intelligence applications and a spectrophotometer, the Easyshade spectrophotometer was found to outperform the Artificial Intelligence applications [28]. To the best of our knowledge, no previous study has evaluated the intra-model repeatability of artificial intelligence chatbots in composite shade selection and their agreement with a dental specialist. Therefore, it is important to investigate the repeatability of artificial intelligence chatbots and their agreement with expert evaluation in composite shade selection performed using standardized photographs. The aim of this study was to evaluate the intra-model repeatability of three artificial intelligence chatbots (ChatGPT-4.0, Copilot, and Claude) in composite shade selection and their agreement with a dental specialist.

The null hypotheses of the study were as follows:

1. Intra-model repeatability in composite shade selection will not show variability among the artificial intelligence chatbots.
2. There will be no agreement between the shade selections of the artificial intelligence chatbots and those of the dental specialist.

2. Materials and Methods

Ethical approval was not required for this research as no human participants or animal materials were involved in the study. The Artificial Intelligence Chatbots used in this study were ChatGPT-4.0 (OpenAI, San Francisco, CA, USA), Microsoft Copilot Pro (Microsoft Corporation, Redmond, WA, USA), and Claude 3.5 Sonnet (Anthropic, San Francisco, CA, USA). All Artificial Intelligence Chatbots were accessed via their paid versions, as free-use plans impose usage and quota limitations that could interfere with repeated evaluations and multi-sample testing required by the study design. Ten acrylic resin maxillary central incisor teeth representing different shades were selected from the VITA Classical shade

guide (VITA Zahnfabrik). The selected shades were A1, A2, A3, B1, B2, C1, C2, D2, D3, and D4.

In clinical practice, shade selection for group-shaded composites (Neo Spectra[®] ST, Dentsply Sirona, Charlotte, NC, USA) typically begins with the clinician selecting the closest match from the VITA Classical shade guide, after which the manufacturer specifies the corresponding composite shade. To replicate this workflow in a highly standardized in vitro setting and to minimize variability arising from the polychromatic nature of natural teeth [2], blinded acrylic resin tooth samples selected from the VITA Classical shade guide were used. The selected samples were chosen based on the manufacturer's shade conversion chart, which indicated that the three main composite shades (A1, A2, and A3) correspond to ten different shades in the VITA Classical system (manufacturer's conversion chart). To ensure blinding, all samples were covered with numbered labels. Customized composite shade tabs were fabricated using a composite resin in shades A1, A2, and A3. The shade selection performance of the Artificial Intelligence Chatbots was evaluated based on photographs containing the blinded sample together with the A1, A2, and A3 composite shade tabs.

To standardize the imaging conditions, a photography product tent equipped with a D65 (6500 K) light source and a gray background was used. Within the tent, the samples were fixed on a surface inclined at 45 degrees. All photographs were taken using a DSLR camera (Canon EOS R5; Canon Inc., Tokyo, Japan) equipped with a macro lens (Canon EF 100 mm; Canon Inc., Tokyo, Japan). The camera was mounted on a tripod and positioned at a fixed distance of 30 cm from the samples. The shooting parameters were set to a shutter speed of 1/200 s, an aperture of f/32, and ISO 160. For each of the ten blinded tooth shades, one standardized photograph was obtained under identical conditions. The acquired images were used directly for artificial intelligence evaluations without any cropping, resizing, or calibration procedures. All photographs were exported in JPEG format, and the same original image file was uploaded to each Artificial Intelligence Chatbot without any modification. All chatbot evaluations were conducted via the web-based interfaces of the respective platforms during April 2025.

Prior to the artificial intelligence assessments, visual shade selection was performed by two Dental Specialists experienced in visual shade matching in restorative dentistry. The evaluations were conducted independently. In cases of disagreement, the samples were re-evaluated and a consensus was reached, resulting in a single expert reference assessment for each sample. Consensus ratings were used as the expert reference standard.

A separate user account was created for each Artificial Intelligence Chatbot. To eliminate the influence of previous responses, a new chat interface was initiated for each evaluation and the chat history was cleared before each session. The following prompt was used: "I will provide you with a photograph showing an acrylic resin tooth in one of the VITA Classical shades together with A1, A2, and A3 composite shade tabs. Can you identify which composite shade most closely matches the acrylic resin tooth?" After the prompt, one photograph was uploaded and the response was recorded. All interactions were conducted in Turkish. The same question and the same ten photographs were used for all models. For each Artificial Intelligence Chatbot, the evaluation procedure was repeated on five different days using new conversation windows each time, and all shade selections were recorded for subsequent analysis. A schematic overview of the study workflow is presented in Figure 1.

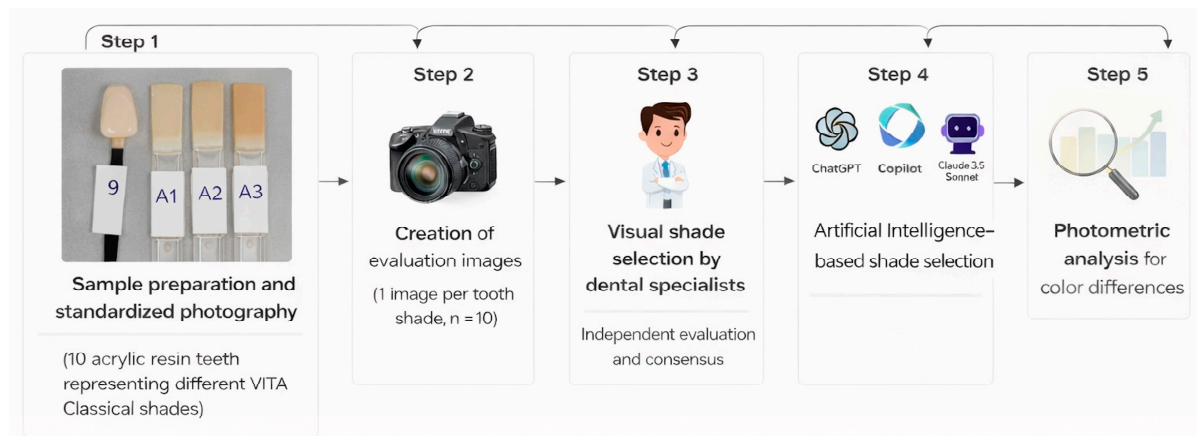


Figure 1. Schematic overview of the experimental workflow used in the present study.

To determine the color differences between the acrylic resin teeth and the composite shade tabs, photometric analysis was performed [29]. Artificial Intelligence Chatbots evaluated non-calibrated JPEG images, whereas photometric ΔE_{00} calculations were performed on calibrated RAW images using a standardized gray reference card. Color measurements were obtained using the same DSLR camera and macro lens, together with a gray reference card (Genuine White Balance Reference; Michael Tapes Design, Inc., New York, NY, USA) and a polarizing filter (Photodental; Dental AG, Inc., Santa Barbara, CA, USA). All images used for photometric analysis were captured in RAW format to avoid compression-related color distortions. The reference color values of the gray card were $L^* = 75$, $a^* = 0$, and $b^* = 0$. The camera was positioned at a fixed distance of 30 cm from the samples. The shooting parameters for color measurements were set to an aperture of $f/18$, a shutter speed of $1/200$ s, and ISO 100. After image acquisition, all photographs were calibrated according to the gray reference card using an image editing software program (Lightroom (<https://lightroom.adobe.com/>); Adobe Inc., San Jose, CA, USA). For each acrylic resin tooth, CIE L^* , a^* , and b^* values were obtained from the center of the middle third region using a digital color measurement software (Digital Color Meter, Version 1.8.1; Apple Inc., Cupertino, CA, USA). Corresponding CIE L^* , a^* , and b^* measurements were obtained from the composite shade tab selected by each evaluator (Artificial Intelligence Chatbot or Dental Specialist) (Table 1). For each evaluator, ΔE_{00} values were calculated based on the composite shade selected by that evaluator (A1, A2, or A3). Therefore, ΔE_{00} values reflect the color difference associated with each evaluator's shade selection. The color differences between the acrylic resin teeth and the composite shade tabs were calculated using the CIEDE2000 formula [30]. ΔE_{00} values were computed in Microsoft Excel (Microsoft, USA) using the CIEDE2000 color difference formula provided by Sharma, with all parametric factors set to 1 [31].

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L'}{k_{LSL}}\right)^2 + \left(\frac{\Delta C'}{k_{CSC}}\right)^2 + \left(\frac{\Delta H'}{k_{HSH}}\right)^2 + R_T \left(\frac{\Delta C'}{k_{CSC}}\right) \left(\frac{\Delta H'}{k_{HSH}}\right)}$$

In the present study, CIEDE2000 perceptibility and acceptability limits were defined according to the 50:50% visual assessment model, with ΔE_{00} values of 0.8 and 1.8 representing the perceptibility and clinical acceptability thresholds, respectively [32]. Statistical analyses were performed using SPSS software (Version 26.0; IBM Corp., Armonk, NY, USA). Intra-model repeatability was assessed using Fleiss' kappa coefficient. Agreement with the dental specialist was calculated separately for each of the five repeated assessments using Cohen's kappa (κ) with 95% confidence intervals. Color difference (ΔE_{00}) values obtained from the

artificial intelligence models and the dental specialist were compared using the Friedman test, followed by post hoc pairwise Wilcoxon signed-rank tests with Holm adjustment when appropriate. A significance level of $p < 0.05$ was considered statistically significant.

Table 1. L*, a*, and b* values of acrylic resin shade-guide teeth from the VITA Classical system and composite shade tabs based on photometric analysis data.

Sample	L*	a*	b*
B2	73.3	2.9	22.6
A1	76.7	1.9	16.9
B1	77.6	1.1	16.5
A2	74.7	3.0	20.7
D4	67.6	3.2	26.7
D3	65.6	7.5	25.3
C1	71.3	2.5	20.2
C2	65.3	4.8	24.1
D2	71.0	2.4	14.9
A3	69.7	4.8	24.8
Composite A1	69.38	2.33	17.03
Composite A2	65.81	4.48	20.81
Composite A3	66.20	6.30	26.20

3. Results

3.1. Repeatability of Artificial Intelligence Models

Intra-model repeatability of artificial intelligence models based on repeated assessments is presented in Table 2. Intra-model repeatability analysis using Fleiss' kappa revealed marked differences among the artificial intelligence models. Claude 3.5 demonstrated moderate repeatability ($\kappa = 0.45$), whereas ChatGPT-4.0 exhibited fair repeatability ($\kappa = 0.33$). In contrast, Microsoft Copilot demonstrated poor repeatability ($\kappa = -0.12$), indicating substantial variability in repeated shade selections.

Table 2. Intra-model reliability of artificial intelligence models based on repeated assessments.

Model	Fleiss' Kappa	95% CI	Interpretation
ChatGPT-4.0	0.327	0.063–0.530	Fair
Microsoft Copilot	−0.119	−0.199–0.035	Poor
Claude 3.5	0.450	0.149–0.702	Moderate

Intra-model repeatability of artificial intelligence models based on five repeated assessments using Fleiss' kappa coefficient. Higher kappa values indicate greater consistency of repeated shade selections.

3.2. Trial-Level Agreement with the Dental Specialist

Agreement between each Artificial Intelligence Chatbot assessment and the dental specialist is presented in Tables 3–5. ChatGPT-4.0 demonstrated variable agreement across repeated assessments, ranging from fair to almost perfect ($\kappa = 0.25$ –0.81). Microsoft Copilot showed consistently slight to poor agreement across assessments ($\kappa = 0.00$ –0.06). Claude 3.5 exhibited fair agreement in some assessments but slight agreement in others, indicating performance variability across sessions.

Across the five repeated assessments (50 total evaluations per model), ChatGPT-4.0 achieved an overall agreement rate of 64%, Claude 3.5 achieved 50%, and Microsoft Copilot achieved 24%. Considering that three composite shade options were available, chance-level agreement would be 33.3%. ChatGPT-4.0 exceeded chance-level performance in all five assessments, Claude 3.5 exceeded chance level in most assessments, whereas Microsoft Copilot's performance approached or fell below chance level in several assessments.

Table 3. Agreement Between ChatGPT-4.0 and the Dental Specialist Across Five Repeated Assessments.

Assessment (1–5)	Agreement (%)	Cohen's κ	95% CI	Agreement Level
1	60.0	0.355	−0.10–0.74	Fair
2	50.0	0.254	−0.18–0.63	Fair
3	90.0	0.808	0.49–1.00	Almost Perfect
4	70.0	0.524	0.10–0.87	Moderate
5	50.0	0.275	−0.15–0.66	Fair

Agreement between each Artificial Intelligence Chatbot assessment and the dental specialist's visual evaluation was calculated using Cohen's kappa (κ) with 95% confidence intervals. The significance level was set at $\alpha = 0.05$.

Table 4. Agreement Between Microsoft Copilot and the Dental Specialist Across Five Repeated Assessments.

Assessment (1–5)	Agreement (%)	Cohen's κ	95% CI	Agreement Level
1	40.0	0.063	−0.33–0.47	Slight
2	30.0	0.028	−0.36–0.42	Slight
3	30.0	0.054	−0.35–0.45	Slight
4	10.0	0.000	−0.29–0.29	Poor
5	10.0	0.000	−0.29–0.29	Poor

Agreement between each Artificial Intelligence Chatbot assessment and the dental specialist's visual evaluation was calculated using Cohen's kappa (κ) with 95% confidence intervals. The significance level was set at $\alpha = 0.05$.

Table 5. Agreement Between Claude 3.5 and the Dental Specialist Across Five Repeated Assessments.

Assessment (1–5)	Agreement (%)	Cohen's κ	95% CI	Agreement Level
1	60.0	0.394	−0.08–0.76	Fair
2	60.0	0.394	−0.08–0.76	Fair
3	30.0	0.014	−0.39–0.40	Slight
4	50.0	0.206	−0.23–0.62	Fair
5	50.0	0.057	−0.30–0.47	Slight

Agreement between each Artificial Intelligence Chatbot assessment and the dental specialist's visual evaluation was calculated using Cohen's kappa (κ) with 95% confidence intervals. The significance level was set at $\alpha = 0.05$.

3.3. Tooth Shade-Specific Agreement (%) with the Dental Specialist Across Five Repeated Assessments

Figure 2 shows the tooth shade-specific agreement percentages between the artificial intelligence models and the dental specialist across five repeated assessments. Agreement percentages differed across tooth shades and among the artificial intelligence models for the same shade.

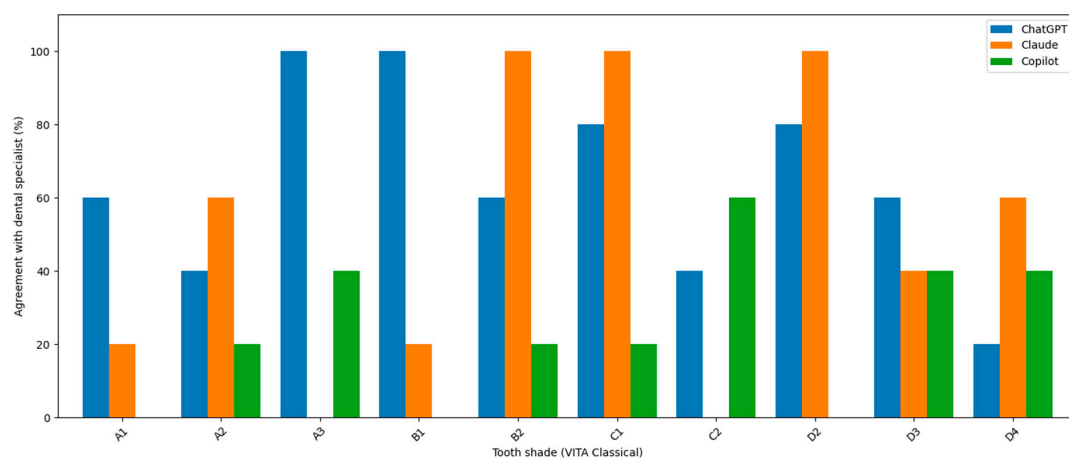


Figure 2. Tooth shade-specific agreement (%) between the artificial intelligence models and the dental specialist across five repeated assessments. Agreement percentages were calculated as the proportion of correct matches between the artificial intelligence outputs and the dental specialist's visual evaluation for each tooth shade.

3.4. Color Difference Comparison Between Artificial Intelligence Models and the Dental Specialist

Table 6 presents the comparison of color differences (ΔE_{00}) between the artificial intelligence models and the dental specialist. The lowest mean ΔE_{00} value was observed for the dental specialist (3.46 ± 1.61), followed by ChatGPT-4.0 (3.96 ± 1.97), Claude 3.5 (4.62 ± 2.76), and Microsoft Copilot (5.31 ± 2.92). ChatGPT-4.0 did not differ significantly from the dental specialist ($p > 0.05$). In contrast, both Claude 3.5 and Microsoft Copilot demonstrated significantly higher ΔE_{00} values compared with the dental specialist ($p < 0.05$). Among the artificial intelligence models, ChatGPT-4.0 exhibited significantly lower ΔE_{00} values than Microsoft Copilot ($p < 0.05$), whereas no statistically significant difference was observed between ChatGPT-4.0 and Claude 3.5 ($p > 0.05$).

Table 6. Comparison of color differences (ΔE_{00}) between artificial intelligence models and the dental specialist.

Group	Mean $\Delta E_{00} \pm SD$
Dental Specialist	3.46 ± 1.61^A
ChatGPT-4.0	3.96 ± 1.97^{AB}
Claude 3.5	4.62 ± 2.76^B
Microsoft Copilot	5.31 ± 2.92^C

Different superscript letters indicate statistically significant differences ($p < 0.05$).

4. Discussion

This study evaluated the intra-model repeatability of three artificial intelligence chatbots in composite shade selection and their agreement with a dental specialist. According to the results, ChatGPT-4.0 demonstrated fair intra-model repeatability, whereas Claude 3.5 showed moderate intra-model repeatability. Therefore, the first null hypothesis, stating that intra-model repeatability would not show variability among the artificial intelligence chatbots, was partially rejected. Trial-level analyses revealed variable agreement between the artificial intelligence chatbots and the dental specialist across repeated assessments. Although ChatGPT-4.0 demonstrated fair to almost perfect agreement based on point estimates, statistically significant agreement beyond chance was observed only in two of the five evaluations, as indicated by confidence intervals not crossing zero. In contrast, for Microsoft Copilot and Claude 3.5, all confidence intervals included zero, indicating that agreement was not statistically significant. Accordingly, the second null hypothesis was not rejected overall; however, statistically significant agreement was observed for ChatGPT-4.0 in two specific assessments.

A previous study also reported that ChatGPT-4.0 provided higher accuracy in dentistry-specific queries [25]. Additionally, ChatGPT-4.0 was found to yield promising results in the Emergency Medicine Board Examination [12]. Mohammad-Rahimi and colleagues reported that ChatGPT-4 achieved the highest average score across various dental disciplines [13]. In the present study, ChatGPT-4.0 did not differ statistically from the dental specialist in terms of ΔE_{00} values. However, it should be noted that the mean color differences of both groups exceeded commonly reported clinical acceptability thresholds.

Another study evaluated chatbot responses to prosthodontics-related questions from a dental specialty board exam, involving five different models: ChatGPT-3.5, Gemini Advanced, Claude Pro, Microsoft Copilot, and Perplexity. Copilot was reported to achieve an accuracy rate of 73% [18]. In contrast, in the present study only slight agreement was observed between Microsoft and the dental specialist, and Microsoft Copilot exhibited significantly higher ΔE_{00} values compared with both the other models and the dental specialist.

In the study by Ozdemir and Yapıcı, ChatGPT-4.0 demonstrated high reliability in restorative dentistry-related questions, despite exhibiting limited consistency [25]. Although ChatGPT-4.0 demonstrated relatively higher agreement with the dental specialist compared with the other artificial intelligence models, this agreement remained only at a fair level, and its intra-model repeatability was also fair. The similarity of its color difference values to those of the dental specialist indicates that the model may generate results close to expert evaluation; however, its stability across repeated assessments appears to be limited. In a study comparing three Artificial Intelligence Chatbots (GPT-4.0, Claude-2, and Llama-2) in answering multiple-choice questions related to prosthodontics and restorative dentistry, GPT-4.0 was reported to demonstrate the highest accuracy and logical quality, whereas the accuracy of Claude-2 was significantly lower compared with GPT-4.0 [19]. In the present study, although Claude 3.5 demonstrated moderate intra-model repeatability, its agreement with the dental specialist remained inconsistent, ranging from fair to slight across assessments, and its color difference values were higher than those of the specialist. These findings suggest that higher repeatability does not necessarily translate into closer agreement with dental specialist evaluation.

Recent dentistry-focused studies have emphasized that the reliability and clinical safety of Large Language Model-based chatbots should be interpreted with caution, particularly in decision-support contexts that may directly influence patient care. Coşkun and Erten Tayşi [26] evaluated the performance of Artificial Intelligence Chatbots in medication-related dental scenarios using standardized prescription prompts in Turkish and reported that newer models demonstrated higher accuracy, completeness, and response quality; however, expert supervision remained essential due to variability in response consistency and the potential for clinically relevant errors. Similarly, Tayeb et al. [27] compared the performance of multiple Artificial Intelligence Chatbots with that of oral surgeons in detecting clinically significant drug–drug interactions and identified substantial differences in sensitivity and specificity among models. Importantly, the authors concluded that although Artificial Intelligence systems may support clinical decision-making, they cannot substitute professional judgment and should be used as supervised decision-support tools rather than autonomous decision-makers. Although these studies primarily focused on text-based, clinically sensitive tasks, their findings align with the present results by highlighting a common limitation of Artificial Intelligence Chatbots: acceptable performance in certain contexts does not necessarily imply stable, repeatable, or expert-level reliability. The present study extends this concern to visually driven esthetic tasks, demonstrating that even under highly standardized photographic conditions, intra-model repeatability and agreement with expert shade selection remain limited. It should be noted that part of the observed day-to-day variability may not solely reflect model inconsistency, but could also be influenced by ongoing updates, non-deterministic behavior, or internal system changes inherent to commercial Artificial Intelligence Chatbots, which are beyond user control. Although the standardized in vitro design allowed controlled comparison among models, it does not fully reproduce intra-oral clinical conditions. Natural teeth exhibit complex surface texture, translucency gradients, surrounding structures, and dynamic lighting conditions that may influence shade perception. In addition, restricting the decision space to three composite shades represents a simplified scenario. Therefore, the present findings should be interpreted as exploratory, and real-world clinical performance may differ.

In the Vita Classical system, tooth shades are grouped into four categories: A, B, C, and D, each representing a dominant hue. Group A includes reddish-brown shades, B includes reddish-yellow, C includes gray, and D includes reddish-gray tones [9]. Shade-specific analysis revealed a pronounced variability in model performance depending on the tooth shade. ChatGPT-4.0 demonstrated higher agreement for certain shades (A3 and B1) but

showed reduced performance for darker shades. Similarly, Claude 3.5 achieved perfect agreement for selected shades while failing in others. These findings suggest that artificial intelligence-assisted shade selection may exhibit a strong shade-dependent behavior rather than uniform performance across all color groups. Browning et al. reported that none of the tested material–shade combinations fell below the clinical acceptability threshold when comparing Vitapan Classical shade tabs with corresponding composite resins [33]. This finding may explain why the color difference values obtained by the dental specialist in the present study exceeded the clinical acceptability limit.

This study has certain limitations. Although the photographs containing VITA Classical shade guide samples and composite shade references were obtained under standardized conditions, they may not fully replicate intra-oral clinical settings. The use of the VITA Classical shade guide for shade selection limits the generalizability of the findings to natural dentition. Additionally, the optical properties of these shade guide samples differ from those of human enamel and dentin, which may influence light reflection and color perception. Furthermore, color difference assessment in the present study was based on photometric analysis derived from calibrated digital photographs; although this approach allows standardized comparisons, it is not fully equivalent to direct spectrophotometric measurements and should be interpreted as an indirect estimation of color difference. Moreover, since only 10 different VITA Classical shades were included in this study, further research is needed to assess the performance of chatbots across a broader range of tooth shades. In addition, the prompts provided to the Artificial Intelligence Chatbots were written in Turkish. Given that large language models are predominantly trained on English-language data, prompt language may have influenced chatbot performance and should be considered a limitation of the present study. Furthermore, only a single standardized prompt formulation was used to evaluate repeatability; future studies may explore the use of alternative prompt structures to assess the robustness of chatbot responses. Future studies may benefit from directly comparing chatbot responses to identical visual prompts presented in Turkish and English to further examine the effect of prompt language on performance. The relatively small sample size ($N = 10$) may have contributed to wide confidence intervals and should be considered when interpreting agreement estimates.

5. Conclusions

Artificial Intelligence Chatbots exhibited variable intra-model repeatability and agreement with expert evaluation in composite shade selection under standardized in vitro conditions. Among the evaluated models, ChatGPT-4.0 demonstrated relatively higher agreement with the dental specialist; however, this agreement was generally within the fair range, despite occasional higher agreement in individual assessments, and its intra-model repeatability remained limited. These findings should be interpreted as exploratory, as controlled laboratory conditions and a constrained decision space may not fully reflect real-world clinical scenarios. Therefore, Artificial Intelligence-assisted shade selection should be considered a potential supportive tool rather than a substitute for professional clinical judgment. Further refinement of visual analysis capabilities and validation under real clinical conditions are required to enhance the reliability and clinical applicability of these systems.

Author Contributions: Conceptualization, S.B.O. and B.O.; methodology, S.B.O., B.O. and C.U.; software, S.B.O.; validation, S.B.O., B.O. and C.U.; formal analysis, S.B.O.; investigation, S.B.O., B.O. and C.U.; resources, S.B.O.; data curation, S.B.O.; writing—original draft preparation, S.B.O. and B.O.; writing—review and editing, S.B.O., B.O. and C.U.; visualization, B.O.; supervision, C.U.; project administration, S.B.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of the current study are listed within the article.

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

References

1. Igiel, C.; Weyhrauch, M.; Wentaschek, S.; Scheller, H.; Lehmann, K.M. Dental color matching: A comparison between visual and instrumental methods. *Dent. Mater. J.* **2016**, *35*, 63–69. [[CrossRef](#)] [[PubMed](#)]
2. Nahsan, F.P.S.; Mondelli, R.F.L.; Franco, E.B.; Naufel, F.S.; Ueda, J.K.; Schmitt, V.L.; Baseggio, W. Clinical strategies for esthetic excellence in anterior tooth restorations: Understanding color and composite resin selection. *J. Appl. Oral Sci.* **2012**, *20*, 151–156. [[CrossRef](#)] [[PubMed](#)]
3. Paolone, G.; Orsini, G.; Manauta, J.; Devoto, W.; Putignano, A. Composite shade guides and color matching. *Int. J. Esthet. Dent.* **2014**, *9*, 164–182.
4. Diamantopoulou, S.; Papazoglou, E.; Margaritis, V.; Kakaboura, A. Change of optical properties of contemporary polychromatic resin composites after light curing and finishing. *Int. J. Esthet. Dent.* **2014**, *9*, 224–237.
5. Diamantopoulou, S.; Papazoglou, E.; Margaritis, V.; Lynch, C.D.; Kakaboura, A. Change of optical properties of contemporary resin composites after one week and one month water ageing. *J. Dent.* **2013**, *41*, e62–e69. [[CrossRef](#)]
6. Vanini, L.; Mangani, F.M. Determination and communication of color using the five color dimensions of teeth. *Pract. Periodontics Aesthetic Dent.* **2001**, *13*, 19–26.
7. Pereira Sanchez, N.; Powers, J.M.; Paravina, R.D. Instrumental and visual evaluation of the color adjustment potential of resin composites. *J. Esthet. Restor. Dent.* **2019**, *31*, 465–470. [[CrossRef](#)]
8. Trifkovic, B.; Powers, J.M.; Paravina, R.D. Color adjustment potential of resin composites. *Clin. Oral Investig.* **2018**, *22*, 1601–1607. [[CrossRef](#)]
9. Brewer, J.D.; Wee, A.; Seghi, R. Advances in color matching. *Dent. Clin.* **2004**, *48*, 341–358. [[CrossRef](#)]
10. Künzle, P.; Paris, S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin. Oral Investig.* **2024**, *28*, 575. [[CrossRef](#)] [[PubMed](#)]
11. Freire, Y.; Laorden, A.S.; Pérez, J.O.; Sánchez, M.G.; García, V.D.-F.; Suárez, A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *J. Prosthet. Dent.* **2024**, *131*, 659.e1–659.e6. [[CrossRef](#)]
12. Liu, C.; Ho, C.; Wu, T. Custom GPTs enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? a study on the emergency medicine specialist examination. *Healthcare* **2024**, *12*, 1726. [[CrossRef](#)]
13. Mohammad-Rahimi, H.; Khoury, Z.H.; Alamdari, M.I.; Rokhshad, R.; Motie, P.; Parsa, A.; Tavares, T.; Sciubba, J.J.; Price, J.B.; Sultan, A.S. Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2024**, *137*, 508–514. [[CrossRef](#)] [[PubMed](#)]
14. Azimi, I.; Qi, M.; Wang, L.; Rahmani, A.M.; Li, Y. Accuracy and Consistency of LLMs in the Registered Dietitian Exam: The Impact of Prompt Engineering and Knowledge Retrieval. *arXiv* **2024**, arXiv:2408.02964. [[CrossRef](#)]
15. Khaldi, A.; Machayekhi, S.; Salvagno, M.; Maniaci, A.; Vaira, L.A.; La Via, L.; Taccone, F.S.; Lechien, J.R. Accuracy of ChatGPT responses on tracheotomy for patient education. *Eur. Arch. Oto-Rhino-Laryngol.* **2024**, *281*, 6167–6172. [[CrossRef](#)]
16. Schwendicke, F.A.; Samek, W.; Krois, J. Artificial intelligence in dentistry: Chances and challenges. *J. Dent. Res.* **2020**, *99*, 769–774. [[CrossRef](#)] [[PubMed](#)]
17. Shan, T.; Tay, F.; Gu, L. Application of artificial intelligence in dentistry. *J. Dent. Res.* **2021**, *100*, 232–244. [[CrossRef](#)]
18. Eraslan, R.; Ayata, M.; Yagci, F.; Albayrak, H. Exploring the potential of artificial intelligence chatbots in prosthodontics education. *BMC Med. Educ.* **2025**, *25*, 321. [[CrossRef](#)] [[PubMed](#)]
19. Chau, R.C.W.; Thu, K.M.; Yu, O.Y.; Hsung, R.T.-C.; Wang, D.C.P.; Man, M.W.H.; Wang, J.J.; Lam, W.Y.H. Evaluation of Chatbot Responses to Text-Based Multiple-Choice Questions in Prosthodontic and Restorative Dentistry. *Dent. J.* **2025**, *13*, 279. [[CrossRef](#)]
20. Zhang, X.; Liang, Y.; Li, W.; Liu, C.; Gu, D.; Sun, W.; Miao, L. Development and evaluation of deep learning for screening dental caries from oral photographs. *Oral Dis.* **2022**, *28*, 173–181. [[CrossRef](#)]
21. Aminoshariae, A.; Kulild, J.; Nagendrababu, V. Artificial intelligence in endodontics: Current applications and future directions. *J. Endod.* **2021**, *47*, 1352–1357. [[CrossRef](#)]
22. Kong, H.-J.; Eom, S.-H.; Yoo, J.-Y.; Lee, J.-H. Identification of 130 dental implant types using ensemble deep learning. *Int. J. Oral Maxillofac. Implant.* **2023**, *38*, 150–156. [[CrossRef](#)]

23. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of large language models such as ChatGPT for dental medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [[CrossRef](#)]
24. Pow, E.H.N.; Suhaimi, S.E.B.; Tan, P.S.W.; Shaharuddin, N.B.; Soo, S.Y.; Said, S.B.M. Accuracy and Reliability of Smartphone Virtual Shade-Matching Technique: An In Vitro Study. *Int. J. Prosthodont.* **2022**, *36*, 331–337.
25. Ozdemir, Z.M.; Yapici, E. Evaluating the Accuracy, Reliability, Consistency, and Readability of Different Large Language Models in Restorative Dentistry. *J. Esthet. Restor. Dent.* **2025**, *37*, 1740–1752. [[CrossRef](#)]
26. Coşkun, Ü.; Tayşi, A.E. The Use of Artificial Intelligence for Medication Support in Dentistry: A Reliability Assessment of Chatbots. *Clin. Exp. Health Sci.* **2025**, *15*, 866–874. [[CrossRef](#)]
27. Tayeb, S.; Barausse, C.; Pellegrino, G.; Sansavini, M.; Pistilli, R.; Felice, P. Comparing Artificial Intelligence (ChatGPT, Gemini, DeepSeek) and Oral Surgeons in Detecting Clinically Relevant Drug–Drug Interactions in Dental Therapy. *Appl. Sci.* **2025**, *15*, 12851. [[CrossRef](#)]
28. Şahin, N.; Kaleli, N.; Ural, Ç. Evaluation of color matching accuracy using artificial intelligence applications and a spectrophotometer: A photometric analysis. *J. Prosthet. Dent.* **2025**, *134*, 1955.e1–1955.e5. [[CrossRef](#)] [[PubMed](#)]
29. Caglar, A.; Yamanel, K.; Gulsahi, K.; Bagis, B.; Özcan, M. Could digital imaging be an alternative for digital colorimeters? *Clin. Oral Investig.* **2010**, *14*, 713–718. [[CrossRef](#)] [[PubMed](#)]
30. Luo, M.R.; Cui, G.; Rigg, B. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Res. Appl.* **2001**, *26*, 340–350. [[CrossRef](#)]
31. Sharma, G.; Wu, W.; Dalal, E.N. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Res. Appl.* **2005**, *30*, 21–30. [[CrossRef](#)]
32. Paravina, R.D.; Ghinea, R.; Herrera, L.J.; Bona, A.D.; Igiel, C.; Linninger, M.; Sakai, M.; Takahashi, H.; Tashkandi, E.; Mar Perez, M.d. Color difference thresholds in dentistry. *J. Esthet. Restor. Dent.* **2015**, *27*, S1–S9. [[CrossRef](#)] [[PubMed](#)]
33. Browning, W.D.; Contreras-Bulnes, R.; Brackett, M.G.; Brackett, W.W. Color differences: Polymerized composite and corresponding Vitapan Classical shade tab. *J. Dent.* **2009**, *37*, e34–e39. [[CrossRef](#)] [[PubMed](#)]

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