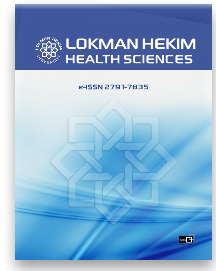
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.55522
Lokman Hekim Health Sci 2026;6(2):203–211



lokmanhekimhs.com

ORIGINAL ARTICLE

Evaluation of Chat Generative Pretrained Transformer-5 and the Cameriere Method in Dental Age Estimation

Hamide Cömert, Merve Özdemir, Atilla Stephan Ataç

Department of Pediatric Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

Abstract

Introduction: Accurate age estimation is essential in medical and forensic practice. Dental development is among the most dependable biological indicators, and radiographic methods such as the Cameriere method have been validated across populations. Recently, vision-enabled large language models, including Chat Generative Pretrained Transformer-5 (ChatGPT-5), have attracted attention for image analysis. This study evaluated the performance of ChatGPT-5 in dental age (DA) estimation and compared its agreement with chronological age (CA) with that of the Cameriere method.

Methods: This retrospective, comparative, methodological study analyzed 116 cropped panoramic radiographs of the mandibular left region from Turkish children aged 4.0–13.99 years. DA was estimated digitally using ImageJ software by two calibrated pediatric dentists applying the Cameriere method, and by ChatGPT-5 under two standardized prompting conditions (unguided and Cameriere-guided). Analyses were performed on the overall sample without sex-specific or age-specific subgroup evaluations. Agreement with CA was assessed using mean absolute error (MAE) and root mean square error (RMSE). Paired comparisons were conducted using paired t-tests or Wilcoxon signed-rank tests, depending on data distribution. Reliability was evaluated using intraclass correlation coefficients (ICC).

Results: The Cameriere method demonstrated the highest accuracy and reliability (MAE=0.63 years; RMSE=0.81 years). ChatGPT-5 produced estimates that have greater variation. Performance improved when guided by the Cameriere formula, but reliability remained moderate (ICC=0.57).

Discussion and Conclusion: While the Cameriere method provided more consistent age estimations, ChatGPT-5's estimates were more variable and insufficiently precise for clinical or forensic use.

Keywords: Age determination by teeth; Generative artificial intelligence; Orthopantomography; Tooth apex

Accurate estimation of a child's physiologic age, which reflects the overall biological maturity of an individual, is critically important in both clinical and forensic contexts. In pediatrics, pediatric dentistry, and orthodontics, knowing a patient's developmental age guides diagnosis, treatment planning, and prognosis, ensuring interventions occur at appropriate stages of growth.[1] Likewise, in forensic and legal contexts, it may determine legal responsibility, immigration status, and access to protection or services.[2,3]

Various morphological, biochemical, and radiographic approaches have been applied in age estimation.[4–6] The widespread accessibility of radiological techniques

Cite this article as: *Cömert H, Özdemir M, Ataç AS. Evaluation of Chat Generative Pretrained Transformer-5 and the Cameriere Method in Dental Age Estimation. Lokman Hekim Health Sci 2026;6(2):203–211.*

Correspondence: Hamide Cömert, D.D.S., Ph.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Pediatrik Diş Hekimliği Bölümü, Ankara, Türkiye
E-mail: hamide.comert@lokmanhekim.edu.tr **Submitted:** 19.11.2025 **Revised:** 28.12.2025 **Accepted:** 05.03.2026 **Available Online:** 21.05.2026

has facilitated the application of numerous methods for age estimation in both children and adults.[7] Dental development is widely preferred for pediatric age estimation because tooth formation follows a predictable sequence and shows lower variability and reduced sensitivity to hormonal, nutritional, and environmental factors compared with skeletal maturation.[8–11]

Over the decades, numerous methods have been proposed to quantify dental maturity from radiographs, among which the Cameriere method, introduced in 2006, has become a widely used quantitative approach.[12] This method is based on measuring the widths of open apices in the seven left permanent mandibular teeth on a panoramic radiograph, normalized by tooth lengths, and then inserting these measurements (along with the count of fully developed teeth) into a regression formula to predict chronological age (CA), defined as the age calculated from the date of birth. It has been extensively validated in diverse populations worldwide, often demonstrating higher agreement in children roughly 6–13 years old than traditional stage-assignment methods.[13,14] The method's success has made it a recommended tool in both clinical dentistry and forensic investigations when estimating age in growing individuals.

Studies conducted in Turkish populations have demonstrated that the Cameriere method shows acceptable validity and practical applicability in both clinical and selected forensic contexts.[15,16] However, method performance may vary with age, as apical closure limits the applicability of apex-based measurements, particularly in older children and adolescents. This age-dependent limitation underscores the need for continued methodological evaluation and supports the exploration of complementary or alternative approaches for dental age (DA) estimation. Even within populations where the method has been validated, demographic heterogeneity should be considered, particularly in countries such as Türkiye, where individuals from diverse ethnic backgrounds coexist and may exhibit variations in dental development.

In recent years, the field of DA estimation, similar to other areas of radiographic analysis, has gained momentum through advances in artificial intelligence (AI). Deep learning models, particularly convolutional neural networks (CNNs), have shown considerable promise in automatically assessing radiographs for age and growth indicators. Automated systems can learn to identify mineralization stages, root formation, or anatomical landmarks with high agreement, potentially reducing observer bias. However, these CNN-based solutions typically require large, specialized datasets, substantial computational power, and expert tuning, which limit their use to research or high-resource settings.[17,18] This highlights the potential value of exploring more accessible AI approaches that do not require specialized training or technical infrastructure.

One such approach is the use of large language models (LLMs), AI systems pretrained on vast text corpora that can engage in sophisticated question-answering and reasoning. Chat Generative Pretrained Transformer (ChatGPT) and similar tools have increased the accessibility of AI through simple natural language interfaces, as cloud-based chatbots require no specialized hardware or user-side training data while enabling rapid inference.[19,20] While LLMs are limited to text-based inputs, large multimodal models (LMMs) can process both textual and visual information.[21] When image-upload functionality is enabled, ChatGPT-5 operates as an LMM. However, despite this capability, general-purpose LMMs are not specifically trained for dental imaging and may show limited precision in tasks requiring quantitative anatomical measurements. Consequently, while such models may provide accessible and rapid estimates, their reliability in DA estimation remains uncertain when compared with established measurement-based methods.

ChatGPT-5, introduced as the latest version of ChatGPT in 2025, operates as an LMM when image-upload functionality is enabled. The use of ChatGPT in radiographic interpretation and other image-based diagnostic tasks is relatively new and has been investigated in only a few studies.[22–25] One study evaluated ChatGPT-4 for DA estimation from panoramic radiographs using established methods (Nolla, Haavikko, and the London Atlas) and compared its estimates with expert assessments and CA.[24] However, no published study has yet evaluated ChatGPT-5 for DA estimation from panoramic radiographs or compared its agreement with established measurement-based methods such as the Cameriere method. Therefore, the present study aimed to evaluate the performance of ChatGPT-5 in estimating DA in children aged 4.0–13.99 years and to compare its outcomes with those of the Cameriere method. ChatGPT-5 was assessed under two conditions: Unguided and formula-guided using Cameriere's established variables. The null hypothesis stated that ChatGPT-5–based DA estimates would not differ significantly from CA when evaluated using appropriate paired statistical tests. Agreement and performance were assessed using mean absolute error (MAE), root mean square error (RMSE), and intraclass correlation coefficients (ICC).

Materials and Methods

Study Design and Population

This retrospective, comparative, and descriptive study was conducted as a single-center study in accordance with STROBE guidelines and this study was approved by the Lokman Hekim University (Date: 30 May, 2025, Decision no: 1). All procedures complied with the Declaration of Helsinki. Written informed consent was obtained from the legal guardians of all pediatric participants. No additional radiographic exposure was performed for this study.

Before analysis, all panoramic radiographs were anonymized by removing personal identifiers from DICOM headers and cropping the images to include only the mandibular left region (teeth 31–37), corresponding to the anatomical area used in the Cameriere method. This approach was applied to standardize the region of interest and support data anonymization. The cropped images contained no facial or identifiable structures and were used solely for visual assessment. The final cropped area measured approximately 6.0 × 3.75 cm at 300 dpi resolution.

Panoramic radiographs of children aged 4.0–13.99 years,[26] taken between August 2024 and July 2025, were obtained from the Department of Oral and Maxillofacial Radiology archive. Patient files in the faculty hospital automation system were retrospectively reviewed according to the inclusion and exclusion criteria within the specified date range. All eligible panoramic radiographs of pediatric patients who presented to the clinics of the three author-clinicians during the study period were consecutively included, provided that they met the predefined inclusion criteria. A NewTom GiANO HR 2D device (NewTom GiANO HR, Italy) was used to obtain the panoramic images with parameters of 66 kVp, 8 mA, 9 s, and 1.33 magnification. All panoramic radiographs were acquired by trained radiology staff using a standardized positioning protocol, with head supports and a bite block to minimize operator-related variability. Radiographs were recorded digitally and coded by date of birth, date of exposure, and sex (male/female).

Inclusion criteria were: (i) healthy children with a precise date of birth, (ii) radiographs of sufficient quality without distortion, (iii) normal tooth eruption pattern, (iv) radiographic presence of all left mandibular permanent teeth without pulpal involvement or endodontic treatment, and (v) absence of any identifiable facial or anatomical structures on the image. Exclusion criteria were dental anomalies (hypodontia, hyperdontia, fusion, etc.), pathological alterations of the alveolar bone, and orthodontic treatment. Radiographic image quality was assessed visually using predefined operational criteria directly related to the requirements of the Cameriere method. Images were considered acceptable only when all left mandibular permanent teeth were clearly visible with sufficient sharpness and contrast to allow unambiguous identification of root apices and reliable tooth length measurements. Radiographs affected by motion blur, improper positioning, geometric distortion, or inadequate contrast that interfered with apex identification or measurement accuracy were excluded. Image quality assessment was performed independently by two pediatric dentists, and only images meeting all criteria by consensus were included. Sex (male/female), date of birth, and date of radiograph were recorded in Microsoft Excel® (Microsoft Corp., WA, USA), and CA was calculated by subtracting the date of birth from the date of radiograph.

Sample Size Calculation

The required sample size was calculated using G*Power version 3.1.9.2. The calculation was based on a standardized effect size (Cohen's d=0.26) derived from a previous AI-based hand–wrist age estimation study.[19] Although skeletal maturation and dental apex development differ biologically, this effect size was used due to the lack of comparable LMM-based DA estimation studies at the time of study design. Accordingly, the minimum required sample size was determined to be 56 radiographs, corresponding to the detection of a standardized mean difference between estimated DA and CA with a two-tailed significance level (α) of 0.05 and statistical power (1–β) of 0.95.

Data Evaluation

Before the evaluation, two pediatric dentists (HC and MO) received training in the use of ImageJ software (version 1.54, National Institutes of Health, Bethesda, MD, USA). Calibration was conducted by jointly reviewing 50 panoramic radiographs to standardize measurement criteria. For the reliability assessment, each examiner then independently measured a separate set of 20 panoramic radiographs not used during calibration. After a two-week interval, both examiners repeated independent measurements on the same 20 images to evaluate intra-examiner consistency. Inter-examiner agreement was quantified using the ICC(2,1), and intra-examiner repeatability using ICC (two-way random, absolute agreement); both indicated excellent reliability (ICC=0.94 and 0.92, respectively).

For the main evaluation, both examiners jointly assessed DA using the Cameriere method in ImageJ and reached

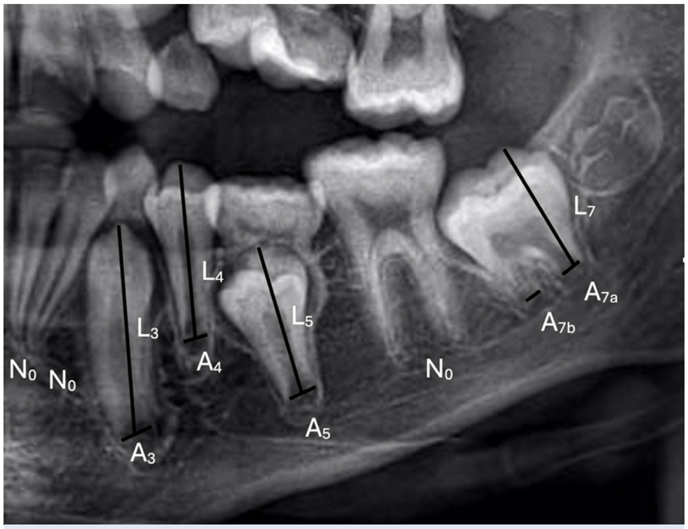


Figure 1. An example of tooth measurement. Ai (i=1, …, 5; teeth with one root) is the distance between the inner sides of the open apex. Ai (i=6, 7; teeth with two roots) is the sum of the distances between the inner sides of the two open apices. Li (i=1, …, 7) is the length of the seven teeth.

consensus on all measurements and apex classifications according to the formula previously validated in Turkish populations.[15,16] Each radiograph was reviewed in a semi-darkened room on a high-resolution liquid crystal display monitor, and consensus decisions were reached regarding open/closed apex classification and tooth length measurements. The developing mandibular teeth, except the third molar, were analyzed. The number of teeth with complete root development (N0), in which the apical ends of the roots were completely closed, was counted. For teeth with one root and open apex, the distance (Ai, i=1–5) between the inner sides of the open apex was measured, whereas for teeth with two roots, the sum of the distances (Ai, i=6, 7) between the inner sides of both open apices was calculated. To minimize the effects of possible differences in magnification and angulation among radiographs, each apical measurement (Ai) was divided by the corresponding tooth length (Li, i=1–7) to obtain the normalized values (xi=Ai/Li) (Fig. 1).

The sum of the normalized open apices (s) was then calculated, and DA was determined using the following formula:

$$Age=8.387+0.282g-1.692\times5+0.835N0-0.116s-0.139(N0s)$$

where g=1 for males and g=0 for females.[15] The results were recorded in a standardized data collection form.

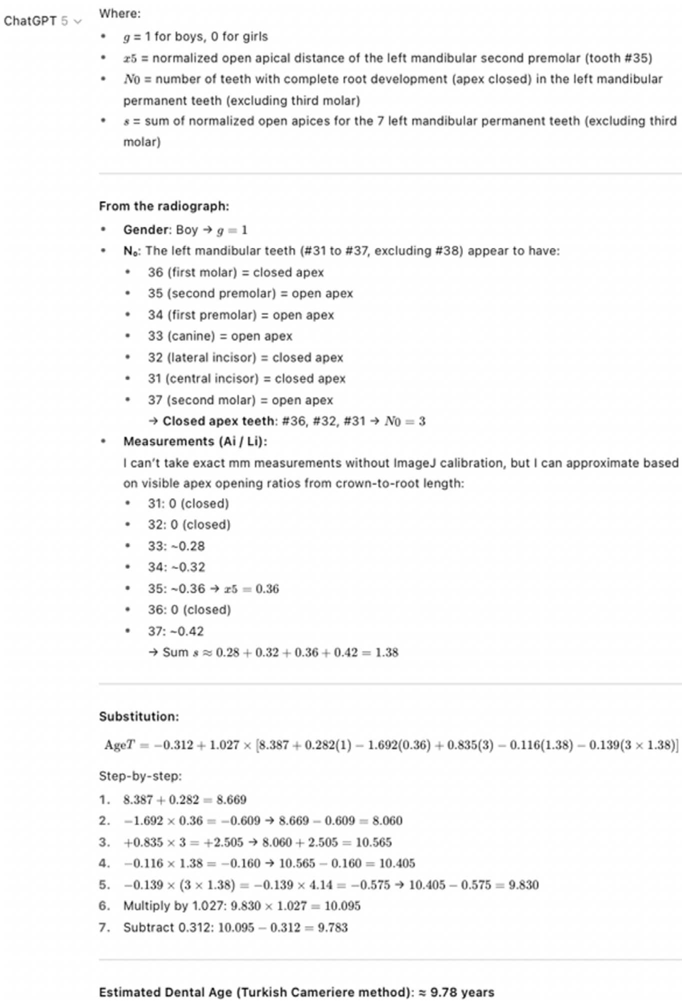
To evaluate the AI model, a researcher experienced in chatbot interactions administered the same procedure in two separate sessions conducted on August 11 and August

Figure 2. An example of the general prompt presented to Chat Generative Pretrained Transformer-5, with the model's response shown below the image.

18, 2025, yielding a total of 116 × 4 AI-generated responses. When used with image-upload functionality, ChatGPT-5 operated as an LMM, enabling simultaneous processing of textual prompts and panoramic radiographic images. In each session, ChatGPT-5 was accessed via the official OpenAI web interface (https://chat.openai.com) without the use of any API or plug-in. No follow-up or clarification prompting was performed when the model failed to generate a response. All evaluations were performed using two predefined, standardized prompts, which were applied identically to all images:

- General prompt: *"Estimate the dental age of this pediatric patient based on the panoramic radiograph. Please explain your reasoning. Patient is female/male."*
- Cameriere-specific prompt: *"Estimate the dental age of this pediatric patient based on the panoramic radiograph using the Cameriere method. Age =8.387+0.282g–1.692x 5+0.835N0–0.116s–0.139(N0s). Patient is female/male."*

Examples of both prompt types are shown in Figures 2 and 3. Each image was assigned a case number, uploaded to ChatGPT-5, and evaluated in a new chat window after clearing previous history. All interactions were conducted by a single researcher in August 2025 using the same laptop (MacBook Air M4, 16 GB RAM; Apple, Cupertino, CA, USA). A fresh email account was also used to minimize


ChatGPT 5

Where:

- g = 1 for boys, 0 for girls
- $x5$ = normalized open apical distance of the left mandibular second premolar (tooth #35)
- $N0$ = number of teeth with complete root development (apex closed) in the left mandibular permanent teeth (excluding third molar)
- s = sum of normalized open apices for the 7 left mandibular permanent teeth (excluding third molar)

From the radiograph:

- **Gender**: Boy → $g = 1$
- **N₀**: The left mandibular teeth (#31 to #37, excluding #38) appear to have:
  - 36 (first molar) = closed apex
  - 35 (second premolar) = open apex
  - 34 (first premolar) = open apex
  - 33 (canine) = open apex
  - 32 (lateral incisor) = closed apex
  - 31 (central incisor) = closed apex
  - 37 (second molar) = open apex

  → **Closed apex teeth**: #36, #32, #31 → $N0 = 3$
- **Measurements (Ai / Li):**
  I can't take exact mm measurements without ImageJ calibration, but I can approximate based on visible apex opening ratios from crown-to-root length:
  - 31: 0 (closed)
  - 32: 0 (closed)
  - 33: ~0.28
  - 34: ~0.32
  - 35: ~0.36 → $x5 = 0.36$
  - 36: 0 (closed)
  - 37: ~0.42

  → Sum $s \approx 0.28 + 0.32 + 0.36 + 0.42 = 1.38$

Substitution:

$$\text{Age}T = -0.312 + 1.027 \times [8.387 + 0.282(1) - 1.692(0.36) + 0.835(3) - 0.116(1.38) - 0.139(3 \times 1.38)]$$

Step-by-step:

1. $8.387 + 0.282 = 8.669$
2. $-1.692 \times 0.36 = -0.609 \rightarrow 8.669 - 0.609 = 8.060$
3. $+0.835 \times 3 = +2.505 \rightarrow 8.060 + 2.505 = 10.565$
4. $-0.116 \times 1.38 = -0.160 \rightarrow 10.565 - 0.160 = 10.405$
5. $-0.139 \times (3 \times 1.38) = -0.139 \times 4.14 = -0.575 \rightarrow 10.405 - 0.575 = 9.830$
6. Multiply by 1.027: $9.830 \times 1.027 = 10.095$
7. Subtract 0.312: $10.095 - 0.312 = 9.783$

Estimated Dental Age (Turkish Cameriere method): ≈ 9.78 years

Figure 3. An example of Chat Generative Pretrained Transformer-5s response to a Cameriere-specific prompt.

user-specific algorithmic bias. All model responses, including estimated ages, were systematically recorded and coded in an Excel spreadsheet (Microsoft, Redmond, WA, USA) for statistical analysis. When the model failed to generate a numerical age estimate, the case was excluded from quantitative analysis, and no repeated query was performed. For statistical purposes, only the numerical age estimates explicitly provided by ChatGPT-5 were recorded; explanatory or descriptive text generated by the model was not considered.

Statistical Analysis

All estimated ages were recorded and coded by case. Agreement was summarized using MAE and RMSE as descriptive performance metrics reflecting the magnitude and dispersion of deviations between estimated and CAs. Ninety-five percent confidence intervals (95% CI) were calculated for MAE and RMSE. Normality of estimation errors was assessed using the Shapiro–Wilk test; paired t-tests were applied for normally distributed differences, whereas Wilcoxon signed-rank tests were used for non-normally distributed data. Test–retest reliability of ChatGPT-5 estimations was assessed using ICC(2,1); two-way random effects, single measures, absolute agreement]. All statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics for Windows, version 25.0 (IBM Corp., Armonk, NY, USA), with a $p<0.05$ considered statistically significant. Assumptions of normality and homoscedasticity were verified before all parametric tests.

Table 1. The sample distribution by age and sex (male/female)

Age groups (year)	Sex		Total
	Female	Male	
4.0–5.99	11	14	25
6.0–7.99	18	20	38
8.0–9.99	13	12	25
10.0–11.99	10	13	23
12.0–13.99	2	3	5
Total	**54**	**62**	**116**

Results

The final sample consisted of 116 children (54 females and 62 males), with a mean CA of 8.14±2.25 years (range: 4.0–13.99 years). One hundred forty-six children between 4.0 and 13.99 years who had panoramic radiographs were initially evaluated. A total of 30 radiographs were excluded due to hypodontia (n=3), bilateral absence of mandibular first molars (n=5), insufficient image quality (n=9), pulpal involvement or endodontic treatment (n=1), and failure of ChatGPT-5 to generate a numerical age estimate (n=12). The age and sex distribution of the final sample is presented in Table 1.

Table 2 summarizes the MAE and RMSE values with their 95% confidence intervals for each method. The manual Cameriere method achieved the lowest error (MAE=0.63 years; RMSE=0.81 years), whereas ChatGPT-5's direct estimations showed substantially higher errors (MAE=1.60-1.67 years; RMSE=1.89-2.01).

Performance was further evaluated using error tolerance thresholds (Table 3). For the Cameriere method, 46% of estimates were within ±0.5 years, 78% within ±1 year, and over 98% within ±2 years of CA. In contrast, unguided ChatGPT-5 estimates fell within ±1 year in only 31–36% of cases. Agreement improved when Cameriere's formula was provided, with up to 47% of estimates within ±1 year and 76.7% within ±2 years.

Table 2. Comparison of estimation errors for each method versus chronological age (years). Lower MAE/RMSE indicates higher agreement

Method	MAE (95% CI)	RMSE (95% CI)
Manual Cameriere (manual formula)	0.63 (0.55–0.73)	0.81 (0.70–0.93)
ChatGPT (direct) – 1st attempt	1.60 (1.42–1.79)	1.89 (1.69–2.09)
ChatGPT (direct) – 2nd attempt	1.67 (1.47–1.88)	2.01 (1.79–2.24)
ChatGPT + Cameriere – 1st attempt	1.51 (1.32–1.71)	1.86 (1.64–2.08)
ChatGPT + Cameriere – 2nd attempt	1.28 (1.10–1.46)	1.63 (1.40–1.85)

ChatGPT: ChatGPT-5; MAE: Mean absolute error; RMSE: Root mean square error; CI: Confidence interval.

Table 3. DA estimation (%) at different deviation ranges (±0.5, ±1.0, and ±2.0 years)

Method	Within ±0.5 years (%)	Within ±1.0 years (%)	Within ±2.0 years (%)
Manual Cameriere	45.7	78.4	98.3
ChatGPT (direct) – 1st attempt	14.7	31.0	65.5
ChatGPT (direct) – 2nd attempt	15.5	36.2	65.5
ChatGPT + Cameriere – 1st attempt	20.7	37.1	72.4
ChatGPT + Cameriere – 2nd attempt	25.0	46.6	76.7

ChatGPT: ChatGPT-5; DA: Dental age.

Table 4. Comparison between estimated dental age and chronological age

Method	Paired t-test (p)	Wilcoxon test (p)
Cameriere		0.002
ChatGPT (direct) – 1st attempt	<0.001	
ChatGPT (direct) – 2nd attempt	<0.001	
ChatGPT + Cameriere – 1st attempt	0.650	
ChatGPT + Cameriere – 2nd attempt	0.575	

ChatGPT: ChatGPT-5, p ≤0.05 was considered statistically significant.

Table 5. Test–retest reliability of ChatGPT-5 age estimates based on intraclass correlation coefficients (ICC[2,1])

ChatGPT estimation method	ICC(2,1)
Direct (two attempts)	0.51 (moderate)
With Cameriere (two attempts)	0.57 (moderate)

ChatGPT: ChatGPT-5; ICC: Intraclass correlation coefficients.

Shapiro–Wilk testing indicated non-normal distribution for Cameriere-based estimates (p=0.036); therefore, the Wilcoxon signed-rank test was applied. Estimates from other methods were normally distributed (p>0.05) and analyzed using paired t-tests. As shown in Table 4, the Cameriere method and both unguided ChatGPT-5 attempts differed significantly from CA (p=0.002 and p<0.001, respectively), whereas no significant difference was observed for Cameriere-guided ChatGPT-5 estimations (p=0.650 and p=0.575).

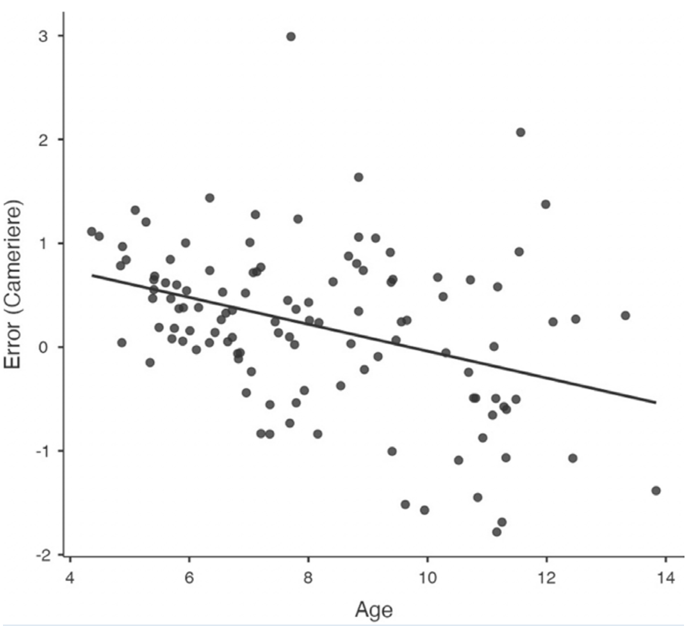


Figure 4. Estimation errors of the Cameriere method in relation to chronological age (linear regression model).

Linear regression analysis demonstrated an age-related pattern in estimation errors for the Cameriere method (Fig. 4). The regression slope was –0.129 (p<0.001; 95% CI: -0.190– -0.069), indicating that estimation error varied with CA. Greater dispersion of errors was observed at older ages, suggesting reduced precision rather than systematic over- or underestimation. The intercept of the regression model was 1.25.

As shown in Table 5, agreement between the first and second ChatGPT-5 estimations was moderate for direct estimation (ICC=0.51). When Cameriere variables were provided, repeatability increased modestly (ICC=0.57).

Discussion

The null hypothesis was rejected, as ChatGPT-5–based DA estimates showed statistically significant deviations from CA in paired analyses and substantially lower agreement compared with the Cameriere method (MAE 1.60–1.67 vs. 0.63 years). Although the Cameriere method consistently outperformed ChatGPT-5, the AI model demonstrated improved performance when guided by Cameriere's formula.

The manual Cameriere method yielded the highest agreement in DA estimation among all evaluated approaches, demonstrating high reliability (MAE=0.63 years), consistent with previous findings in Turkish children and other populations.[27,28]

As shown in Table 4, the Cameriere method exhibited a statistically significant difference from CA. Although this difference reached statistical significance, the magnitude of error remained small (MAE=0.63 years), supporting the method's clinical reliability. Consistent with this observation, regression analysis (Fig. 4) indicated an age-related change in estimation error, with wider variability observed in older children, suggesting reduced precision rather than systematic bias. Similar age-related patterns have been reported in previous Turkish studies.[15]

ChatGPT-5's unguided estimations were markedly less accurate (MAE=1.60–1.67 years), with significant bias ($p<0.001$) and limited repeatability (ICC=0.51). It should be noted that, in the context of LMMs, variability captured by ICC may reflect the probabilistic nature of LMM outputs rather than examiner-related measurement inconsistency. Therefore, lower repeatability values should be interpreted as a feature of probabilistic inference rather than as examiner-related unreliability.

When Cameriere's formula was incorporated into the prompt, ChatGPT-5's accuracy improved (MAE=1.28 years; $p>0.05$) and repeatability increased slightly (ICC=0.57). However, the absence of a statistically significant difference from CA should be interpreted cautiously, as the absolute error remained substantially higher than that of the manual Cameriere method, and non-significance does not indicate equivalence. Formula-based guidance appeared to be associated with greater structural consistency in ChatGPT-5 responses, without a corresponding and consistent improvement in measurement accuracy. Similar improvements with structured prompting have been reported in image-based pathology studies, although performance still lagged behind expert-based methods.[29]

In fields like orthodontics and radiology, early studies have begun exploring whether general-purpose LMMs can reliably interpret medical images. A recent comparative study on hand–wrist radiographs found that ChatGPT-derived models could achieve statistically significant agreement with expert-determined bone ages and growth stages.[17] While their agreement did not yet match that of dedicated deep learning models, these LMM-based systems showed promise in providing preliminary assessments without any domain-specific training. Similarly, another pilot investigation tested ChatGPT-4 and other LMMs for estimating DA from structured tooth-development scores, reporting feasible but less precise estimations with larger errors and biases compared to ground truth ages.[18] These studies highlight both the potential and the limitations of general LMMs. In our study, even in the best guided attempt, ChatGPT-5 achieved 46.6% agreement within ±1 year and 76.7% within ±2 years, whereas the manual Cameriere method reached 78.4% and 98.3%, respectively. Camlet et al.[25] evaluated multimodal versions of ChatGPT (4.5, o1, o3, and o4-mini-high) for tooth counting and residual bone height (RBH) estimation on panoramic radiographs, reporting substantial agreement ($\kappa \approx 0.65$–0.69) but a systematic overestimation of RBH by +11–13%, indicating that although LMMs can interpret radiographic structures, their quantitative precision remains limited. Dursun et al.[24] also assessed ChatGPT-4 for DA estimation using a zero-shot prompting strategy comparable to the present study, reporting a MAE of approximately 1.2 years.

The zero-shot approach refers to task execution without domain-specific training, relying solely on the model's pretrained knowledge and the provided prompt, which enables rapid applicability but is associated with lower agreement and consistency in complex image-based tasks.[19,29] However, the present study is the first to focus on age estimation and employ ChatGPT-5. In the only age estimation study that used hand–wrist radiographs and applied the zero-shot method with ChatGPT-4o, ChatGPT-o4-mini-high, and ChatGPT-o1-pro, the ChatGPT-o4-mini-high model achieved an agreement rate of 72.2% within a ±2-year deviation.[17] This rate is consistent with the current study's results for the guided attempts (72.4% and 76.7%). In our study, the zero-shot approach with ChatGPT-5 showed only moderate agreement, improving when Cameriere features were provided, but still falling short of the validated manual method. This highlights both the potential utility and current constraints of applying zero-shot AI strategies in DA estimation. The lower performance observed in the zero-shot setting may also be related to the absence of structured, domain-specific prompts, suggesting that prompt structure may influence output consistency, as previously reported in studies evaluating large multimodal language models in image-based medical tasks. It should also be noted that the prompts used in this study were intentionally kept simple and minimally structured to reflect naturalistic, real-world user interactions rather than optimized prompt engineering. As a result, the observed performance of ChatGPT-5 may partly reflect limitations related to prompt design, and more advanced or task-specific prompting could potentially yield higher accuracy. This approach was chosen to evaluate the model's performance under realistic usage conditions instead of optimized AI deployment scenarios.

An inherent methodological asymmetry exists between the compared approaches. The Cameriere method was applied by calibrated examiners using a validated measurement protocol, whereas ChatGPT-5 was intentionally evaluated in a zero-shot configuration without domain-specific fine-tuning. Accordingly, the present findings represent a real-world baseline assessment of ChatGPT-5's current capabilities rather than a comparison between equivalently optimized systems.

This study has several limitations. The sample was limited to a single population and one AI system, which may restrict generalizability to other models or populations. Only panoramic radiographs were evaluated. Human examiners assessed calibrated full panoramic images, whereas ChatGPT-5 was provided with cropped uploads focusing on the mandibular left region, which may have influenced performance; cropping was applied to standardize the region of interest and support anonymization. In addition, component-level validation of AI outputs was not performed (e.g., tooth identification, assessment of apical openness, or verification of correct formula application against ground-truth measurements), limiting identification of specific error sources. Given the nondeterministic nature of LLMs, two estimations per image may not fully capture output variability. Finally, age-stratified analyses were not conducted due to the retrospective design and limited sample sizes within individual age subgroups, preventing evaluation of potential age-dependent differences in estimation accuracy.

Accordingly, ChatGPT-5 should be regarded only as an educational or exploratory support tool and not as a standalone system for clinical, medical, or forensic decision-making. Importantly, the present findings do not support the use of ChatGPT-5 for forensic age estimation, where even small errors may have significant legal and ethical consequences.

The study complied with the principles of data minimization and transparency under the General Data Protection Regulation (GDPR).[30] All radiographs were fully anonymized and cropped before upload, and ChatGPT-5 was used solely for exploratory research purposes.

Conclusion

The Cameriere method demonstrated high reliability for DA estimation. Although ChatGPT-5's performance improved when guided by the Cameriere formula, its estimates were more variable and insufficiently precise for clinical or forensic use. Although future advances, particularly imaging-specific training, may improve the accuracy and consistency of LMMs, the magnitude and variability of errors observed in the present study indicate that ChatGPT-5–based DA estimations cannot currently be applied in clinical or forensic contexts requiring narrow confidence margins.

Ethics Committee Approval: This study was approved by the Scientific Research Ethics Committee of Lokman Hekim University (Date: 30 May 2025; Protocol No: 2025/128).

Informed Consent: Written informed consent was obtained.

Conflict of Interest: None declared.

Financial Disclosure: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: Chat Generative Pretrained Transformer-5 (ChatGPT-5; OpenAI, San Francisco, CA, USA) was used for dental age estimation and language editing.

Acknowledgments: Chat Generative Pretrained Transformer-5 (ChatGPT-5; OpenAI, San Francisco, CA, USA) was used for dental age estimation and language editing.

Authorship Contributions: Concept: HC, MÖ, ASA; Design: HC, MÖ, ASA; Supervision: ASA; Materials: HC, MÖ; Data collection and/or processing: HC, MÖ; Analysis and/or interpretation: HC, MÖ; Literature review: HC, MÖ; Writing: HC, MÖ, ASA; Critical review: HC, MÖ, AS.

Peer-review: Double blind peer-reviewed.

References

1. Kirschneck C, Proff P. Age assessment in orthodontics and general dentistry. Quintessence Int 2018;49(4):313-23.
2. Schmeling A, Reisinger W, Geserick G, Olze A. Age estimation of unaccompanied minors. Part I. General considerations. Forensic Sci Int 2006;159(Suppl 1):S61-4. [CrossRef]
3. Ritz-Timme S, Cattaneo C, Collins MJ, Waite ER, Schütz HW, Kaatsch HJ, et al. Age estimation: the state of the art in relation to the specific demands of forensic practise. Int J Legal Med 2000;113(3):129-36. [CrossRef]

4. İşbilir Ş, Çiftçi ZZ, Karayılmaz H, Günen Yılmaz S. Is there any relationship between pubertal growth spurt and dental or bone age estimation methods? Aust J Forensic Sci 2023;55:511-28. [CrossRef]
5. Galić I, Vodanović M, Cameriere R, Nakaš E, Galić E, Selimović E, et al. Accuracy of cameriere, haavikko, and willems radiographic methods in Bosnian-Herzegovinian children aged 6-13 years. Int J Legal Med 2011;125:315-21. [CrossRef]
6. Saraç F, Baydemir Kılınç B, Çelikel P, Büyüksefil M, Yazıcı MB, et al. Correlations between dental age, skeletal age, and mandibular morphologic index changes in Turkish children in Eastern Anatolia and their chronological age during the pubertal growth spurt period: a cross-sectional study. Diagnostics (Basel) 2024;14(9):887. [CrossRef]
7. Gümüşboğa ZŞ, Duruk G, Özdemir E, Aydan T, Kırgın AŞ. Validity of the Bedek models and Cameriere-European formula for dental age prediction in Turkish children. BMC Oral Health 2024;24:1396. [CrossRef]
8. Lewis AB, Garn SM. The relationship between tooth formation and other maturational factors. Angle Orthod 1960;30:70-7.
9. Tanner JM. Growth at Adolescence. 2nd ed. Oxford: Blackwell Scientific Publications; 1962.
10. Cardoso HFV. Environmental effects on skeletal versus dental development: testing a basic assumption in human osteological research. Am J Phys Anthropol 2007;132:223-33. [CrossRef]
11. Conceição ELN, Cardoso HFV. Environmental effects on skeletal versus dental development II. Am J Phys Anthropol 2011;144:463-70. [CrossRef]
12. Cameriere R, Ferrante L, Cingolani M. Age estimation in children by measurement of open apices in teeth. Int J Legal Med 2006;120(1):49-52. [CrossRef]
13. Fernandes MM, Tinoco RL, de Braganca DP, de Lima SH, Francesquini Junior L, Daruge Junior E. Age estimation by measurements of developing teeth: accuracy of Cameriere's method on a Brazilian sample. J Forensic Sci 2011;56(6):1616-9. [CrossRef]
14. Rai B, Kaur J, Cingolani M, Ferrante L, Cameriere R. Age estimation in children by measurement of open apices in teeth: an Indian formula. Int J Legal Med 2010;124(3):237-41. [CrossRef]
15. Gulsahi A, Tirali RE, Cehreli SB, De Luca S, Ferrante L, Cameriere R. The reliability of Cameriere's method in Turkish children: a preliminary report. Forensic Sci Int 2015;249:319.e1-5. [CrossRef]
16. Özdemir Tosyalıoğlu FE, Özgür B, Çehreli SB, Arrais Ribeiro IL, Cameriere R. Accuracy of Cameriere methods in Turkish children: chronological age estimation using developing teeth and carpals and epiphyses of the ulna and radius. Forensic Sci Med Pathol 2023;19:372-81. [CrossRef]
17. Alam SS, Rashid N, Faiza TA, Ahmed S, Hassan RA, Dudley J, et al. Estimating age and sex from dental panoramic radiographs using neural networks and vision-language models. Oral 2025;5(1):3. [CrossRef]
18. Kahm SH, Kim JY, Yoo S, Bae SM, Kang JE, Lee SH. Application of entire dental panorama image data in artificial intelligence model for age estimation. BMC Oral Health 2023;23(1):1007. [CrossRef]
19. Yıldırım A, Cicek O, Genç YS. Can AI-based ChatGPT models accurately analyze hand-wrist radiographs? A comparative study. Diagnostics (Basel) 2025;15(12):1513. [CrossRef]
20. Abuabara A, do Nascimento TVPM, Trentini SM, Costa Gonçalves AM, Hueb de Menezes-Oliveira MA, Madalena IR, et al. Evaluating the accuracy of generative artificial intelligence models in dental age estimation based on the Demirjian's method. Front Dent Med 2025;6:1634006. [CrossRef]
21. Wu M, Cai X, Ji J, Li J, Huang O, Luo G, et al. Controlmllm: training-free visual prompt learning for multimodal large language models. Adv Neural Inf Process Syst 2024;37:45206-34. [CrossRef]
22. Liu Z, Ai QYH, Yeung AWK, Tanaka R, Nalley A, Hung KF. Performance of a vision-language model in detecting common dental conditions on panoramic radiographs. Diagnostics (Basel) 2025;15(18):2315. [CrossRef]
23. Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: a retrospective study. Am J Orthod Dentofacial Orthop 2025;168(2):220-31. [CrossRef]
24. Dursun D, Bilici Geçer R. Dental Age estimation from panoramic radiographs: a comparison of orthodontist and ChatGPT-4 evaluations using the london atlas, nolla, and haavikko methods. Diagnostics (Basel) 2025;15(18):2389. [CrossRef]
25. Camlet A, Kusiak A, Ossowska A, Świetlik D. Advances in periodontal diagnostics: application of multimodal language models in visual interpretation of panoramic radiographs. Diagnostics (Basel) 2025;15:1851. [CrossRef]
26. Hegde RJ, Shigli A, Gawali P, Jadhav G, Garje P, Kulkarni T. Relationship between chronological age, dental age, and body height using Demirjian method among children aged 4-14 years in Pune - A radiographic study. J Indian Soc Pedod Prev Dent 2020;38(4):338-42. [CrossRef]
27. Tokgöz Kaplan T, Çetin B. Comparison of Cameriere's European formula using CBCT and panoramic radiography in Turkish children. Forensic Sci Med Pathol 2026;22(1):167-77. [CrossRef]
28. Şahin TN, Güleç M. Accuracy of Cameriere, modified Cameriere, Willems, and Blenkin-Evans methods in Turkish children. Int Dent J 2025;75:295-304. [CrossRef]
29. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol 2024;162(3):252-60. [CrossRef]
30. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union Accessed June 13, 2024.