ORIGINAL ARTICLE

DOI: 10.4274/hamidiyemedj.galenos.2026.82612

Hamidiye Med J

Classification of Aortic Arch Anatomy into Normal, Variant, and Pathological Categories: Evaluation of a Large Language Model's Classification accuracy

Arcus Aortae Anatomisinin Normal, Varyant ve Patolojik Kategorilere Sınıflandırılması: Bir Büyük Dil Modelinin Sınıflandırma Doğruluğunun Değerlendirilmesi

İsmail Sivri, Furkan Mehmet Özden, Gamze Gül, Emre Kaygın, Tuncay Çolak

Kocaeli University Faculty of Medicine, Department of Anatomy, Kocaeli, Türkiye

ABSTRACT

Background: Large language models are increasingly integrated into medical education and diagnostic support, yet their capacity to interpret anatomically complex, variation-rich vascular structures remains underexplored. This study evaluated the performance of ChatGPT-5.1 Plus in classifying aortic arch configurations as normal, variant, or pathological.

Materials and Methods: Thirty-nine text-based anatomical descriptions derived from authoritative references were used; each description was presented in an independent session with a standardized few-shot prompt requiring structured output. The procedure was repeated after one week to assess temporal consistency. Accuracy was measured across different categories, and statistical agreement was calculated using Cohen's kappa.

Results: The model demonstrated excellent accuracy in identifying normal anatomy (100%) and high performance in recognizing anatomical variants (91.7%). However, pathological patterns were classified with lower accuracy (70.6%) because anomalies were misinterpreted as benign variations. Overall agreement with the reference standard was substantial (Kappa = 0.766; $p < 0.001$). In a binary evaluation comparing normal or variant anatomy with pathological cases, the model achieved an accuracy of 84.62%, a sensitivity of 70.59%, and a specificity of 95.45%. Five of the six were undercalls of pathology.

Conclusion: ChatGPT-5.1 Plus shows significant promise in recognizing standard aortic arch anatomy and common variants. However, its reduced sensitivity to atypical or clinically significant anomalies underscores the necessity for expert oversight when it is used for educational or classification purposes in specialized medical contexts.

Keywords: Anatomical variation, aortic arch, large language model, classification accuracy

ÖZ

Amaç: Büyük dil modelleri giderek tıp eğitimi ve tanısal destek süreçlerine entegre edilmektedir; ancak anatomik açıdan karmaşık ve varyasyon açısından zengin vasküler yapıların yorumlanmasındaki kapasiteleri yeterince araştırılmamıştır. Bu çalışmada, ChatGPT-5.1 Plus modelinin arcus aortae konfigürasyonlarını normal, varyant veya patolojik olarak sınıflandırmadaki performansı değerlendirilmiştir.

Gereç ve Yöntemler: Yetkin kaynaklardan elde edilen 39 adet metin tabanlı anatomik tanım kullanılmıştır. Her bir örnek, yapılandırılmış çıktı gerektiren standartlaştırılmış birkaç örnekli yönlendirme ile bağımsız oturumlarda modele sunulmuştur. Zamansal tutarlılığı değerlendirmek amacıyla aynı prosedür bir hafta sonra tekrarlanmıştır. Doğruluk oranları farklı kategoriler için hesaplanmış, referans standart ile uyum Cohen's kappa katsayısı ile analiz edilmiştir.

Bulgular: Model, normal anatomiyi tanımlamada mükemmel doğruluk (%100) ve anatomik varyantları belirlemede yüksek performans (%91,7) göstermiştir. Buna karşın, patolojik durumların sınıflandırılmasında doğruluk daha düşük bulunmuştur (%70,6); bunun temel nedeni anomalilerin benign varyasyonlar olarak yanlış yorumlanmasıdır. Referans standart ile genel uyum anlamlı

Address for Correspondence: Asst. Prof. İsmail Sivri, MD, Kocaeli University Faculty of Medicine, Department of Anatomy, Kocaeli, Türkiye

E-mail: ismail.sivri@kocaeli.edu.tr **ORCID ID:** orcid.org/0000-0002-5809-5693

Received: 10.04.2026 **Accepted:** 07.09.2026 **Epub:** 30.10.2026

Cite this article as: Sivri İ, Özden FM, Gül G, Kaygın E, Çolak T. Classification of aortic arch anatomy into normal, variant, and pathological categories: evaluation of a large language model's classification accuracy. Hamidiye Med J. [Epub Ahead of Print]

Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of University of Health Sciences Türkiye, Hamidiye Faculty of Medicine. This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

ÖZ

ve yüksek düzeydedir (Kappa = 0,766; p < 0,001). Normal/varyant ile patolojik durumların karşılaştırıldığı ikili analizde, modelin doğruluğu %84,62; duyarlılığı %70,59 ve özgüllüğü %95,45 olarak hesaplanmıştır. Toplam altı hatalı sınıflandırmanın beşinde patolojinin olduğundan daha hafif değerlendirilmesi söz konusudur.

Sonuç: ChatGPT-5.1 Plus, standart aortik ark anatomisini ve yaygın varyantları tanımada önemli bir potansiyel göstermektedir. Bununla birlikte, atipik veya klinik açıdan anlamlı anomalilere karşı duyarlılığının görece düşük olması, özellikle uzmanlık gerektiren tıbbi bağlamlarda eğitimsel veya sınıflandırma amaçlı kullanımlarda uzman denetiminin gerekliliğini ortaya koymaktadır.

Anahtar Kelimeler: Anatomik varyasyon, arcus aortae, büyük dil modeli, sınıflandırma doğruluğu

Introduction

The aortic arch develops from six pairs of pharyngeal arch arteries, which appear between the third and eighth weeks of gestation. These arteries undergo selective regression and remodeling to form the mature great vessels. Disruptions in this process can lead to anatomical variants or congenital anomalies, such as a double aortic arch or a right-sided aortic arch (1).

Variations in the branching pattern of the aortic arch present significant anatomical and clinical challenges. Abnormal origins or unusual configurations of the arch vessels can complicate diagnostic imaging and endovascular procedures. These variations demand meticulous preoperative evaluation and adaptive intraoperative strategies to ensure procedural safety and effectiveness. A detailed understanding of aortic arch morphology, therefore, plays a crucial role in preventing vascular injuries and optimizing surgical outcomes (2).

Today, artificial intelligence (AI) has become an integral part of the medical field, particularly in medical research and clinical practice. While the underlying GPT-3 model was initially announced in June 2020, the release of ChatGPT (version 3.5) in November 2022 laid the groundwork for its use in AI-assisted medical applications. As a large language model (LLM), it has considerable potential to provide medical support. However, various challenges and concerns regarding its implementation in healthcare persist (3). Several studies have investigated ChatGPT's ability to serve as a virtual physician for diagnostic purposes (3-5).

This study aimed to evaluate ChatGPT's ability to recognize and describe variations in the branching pattern of the aortic arch. By analyzing its anatomical accuracy and consistency, we sought to determine whether the model demonstrates sufficient understanding of clinically relevant vascular variations to function as a supportive tool in medical education and research.

Materials and Methods

To ensure objective classification, aortic arch configurations were operationally defined based on established anatomical and clinical criteria (1,6,7). Normal anatomy was defined as the standard three-branch adult configuration originating sequentially from the aortic arch, comprising the brachiocephalic trunk, left common carotid artery, and left subclavian artery (6). Anatomical variants were defined as non-pathological, benign structural or branching deviations resulting from altered embryological remodeling that typically remain asymptomatic and do not impair hemodynamics or normal vascular function, such as a common origin of the brachiocephalic and left common carotid arteries or a persistent left superior vena cava (1,7). Conversely, pathological anomalies were defined as structural malformations or severe congenital aberrations involving inherent developmental defects, hemodynamic compromise, vascular obstruction, or clinically significant symptomatic manifestations, including severe aortic coarctation, interrupted aortic arch, or double aortic arch causing airway compression (1,7).

In this study, 39 text-based anatomical cases were developed using two widely accepted reference sources in anatomy and anatomical variation (6,7). Specifically, descriptions of anatomical variants and pathological configurations were adapted and synthesized from Bergman's Comprehensive Encyclopedia of Human Anatomic Variation (7), while sentences representing normal anatomy were derived from the normoanatomical descriptions in Gray's Anatomy (6). The model was accessed via the official web interface: a standard ChatGPT-5.1 account was created and subsequently upgraded to a ChatGPT-5.1 Plus subscription.

The sentences concerning the aortic arch were presented to ChatGPT-5.1 Plus to evaluate its performance in interpreting and classifying text-based anatomical descriptions, and to assess the temporal consistency of its responses. To ensure unbiased interpretation, the

sentences were carefully constructed to avoid explicit diagnostic cues or lexical hints. Terms such as normal, variant, variation, anomalous, and embryological defect were intentionally excluded, forcing the model to infer the correct category solely on anatomical grounds. To eliminate memory retention bias, each question was asked in a separate temporary chat session within the standard web interface, ensuring no contextual carryover between responses. A randomized ordering of the numbers 1 to 39 was generated using an online randomization tool, and the questions were presented to the model in this sequence (8). To assess temporal stability, the same set of questions was administered at two time points, with the first session conducted on November 15, 2025, and the second session conducted on November 22, 2025. The first session was designated a priori as the primary performance evaluation. The second session was used exclusively for test–retest analysis.

A standardized few-shot prompt that included one example for each classification category (normal, variant, and pathological). Before each test item, the prompt assigned the model the role of an anatomist with 20 years' experience and provided one reference example for each classification category. The normal example described the origin of the internal carotid artery from the common carotid artery at the C4 vertebral level; the variant, a persistent left superior vena cava draining into the coronary sinus; and the pathological, a single arterial trunk resulting from failure of truncoconal septation. These examples were used solely to define the intended classification boundaries and the required response format. None of the few-shot examples was included among the 39 test items, and no test item was an exact or near-verbatim reproduction of an example. With this approach, we aimed to minimize the risks of overfitting and of prompt leakage from the few-shot examples. The complete prompt template with three examples and all test items are provided in Supplementary Files 1 and 2.

Each question required the model to generate a structured JavaScript Object Notation (JSON) output with three fields: (1) description, (2) classification, and (3) reasoning. The description field restated the anatomical configuration to ensure alignment between the question and the model's interpretation. The classification field indicated the model's decision among normal, variant, or pathological categories. Finally, the reasoning field provided a concise anatomical or embryological justification, confirming that the response was based on logical understanding rather than random association.

For the purposes of this study, the categories were operationally defined according to the selected reference sources. Configurations representing the expected adult anatomical arrangement were classified as normal. Recognized structural deviations that lacked inherent pathological conditions were classified as variants, whereas congenital malformations or configurations described as developmentally or clinically abnormal were classified as pathological. The reference classifications were based on the terminology and categorization provided in the selected source materials.

To assess the model's classification accuracy, all 39 text-based items were evaluated within a trichotomous framework comprising normal anatomy, anatomical variants, and pathological (anomalous) configurations. Model outputs were directly compared with the reference standard for each category, and overall performance patterns were summarized using a multiclass confusion matrix. The workflow of this study is presented in Figure 1.

To further evaluate the model's classification accuracy in identifying pathological patterns, predictions were examined using a 2×2 contingency table. Cases labeled as pathological by both the model and the reference standard were defined as true positives (TPs), whereas normal or variant anatomy correctly identified as non-pathological constituted true negatives (TN). Sensitivity (recall), specificity, positive predictive value (PPV; precision), negative predictive value (NPV), accuracy, and F1 score were then calculated from the corresponding true-positive, false-positive, false-negative, and true-negative counts. All metrics were expressed as percentages to facilitate clearer comparisons. The workflow diagram and the misclassification distribution visualizations were generated using Microsoft PowerPoint 365 (Redmond, WA, USA).

This experimental setup evaluates the performance of ChatGPT-5.1 Plus using a specific prompt engineering strategy (combining role assignment, few-shot prompting, and structured JSON output) rather than assessing the intrinsic capabilities of the unprompted baseline model.

This article does not contain any studies with human participants or animals. Ethical approval was not required for this type of publication. Informed consent was not required because the article does not contain any studies with human participants or animals.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 31.0 (Armonk, NY: IBM Corp.). The total and category-specific accuracies (normal, variant, and pathological) of ChatGPT-5.1 Plus were presented as counts and percentages. To assess the agreement between ChatGPT's outputs and the reference answers, Cohen's kappa coefficient for nominal data was calculated at both the overall and subgroup levels. Differences between ChatGPT's

responses and the reference answers across the three categories were evaluated using the McNemar–Bowker test. Temporal differences between the two evaluation sessions were assessed at the distribution level using the McNemar–Bowker test, while item-level reproducibility and test–retest agreement were evaluated using Cohen's Kappa and observed percent agreement. A p value of < 0.05 was considered statistically significant.

For classification accuracy calculations, the three-category system was reduced to two groups. Normal and variant anatomical configurations were combined into the non-pathological group, while pathological configurations were classified as the pathological group. Each performance metric was derived from the counts of TPs, false positives (FPs), false negatives (FNs), and TNs. Sensitivity (Recall) was calculated as TP/(TP + FN), representing the proportion of correctly identified positive cases. Specificity was calculated as TN/(TN + FP). PPV (precision) was computed as TP/(TP + FP). The NPV was calculated as TN/(TN + FN). Overall accuracy was calculated as (TP + TN)/(TP + FP + FN + TN). The F1 score was calculated as 2TP/(2TP + FP + FN). The heatmap visualizations were produced with Microsoft Excel 365 (Redmond, WA, USA). Confidence intervals were calculated using the Clopper-Pearson exact method and the non-parametric percentile bootstrap method.

Results

The model demonstrated high accuracy in identifying normal anatomy (100%) and variant configurations (91.7%), but performance was lower for pathological cases (70.6%), primarily because pathological patterns were misclassified as variants (Table 1). According to the McNemar–Bowker test, no statistically significant difference was observed between the groups (p > 0.05). Figure 2 also presents a heatmap visualization.

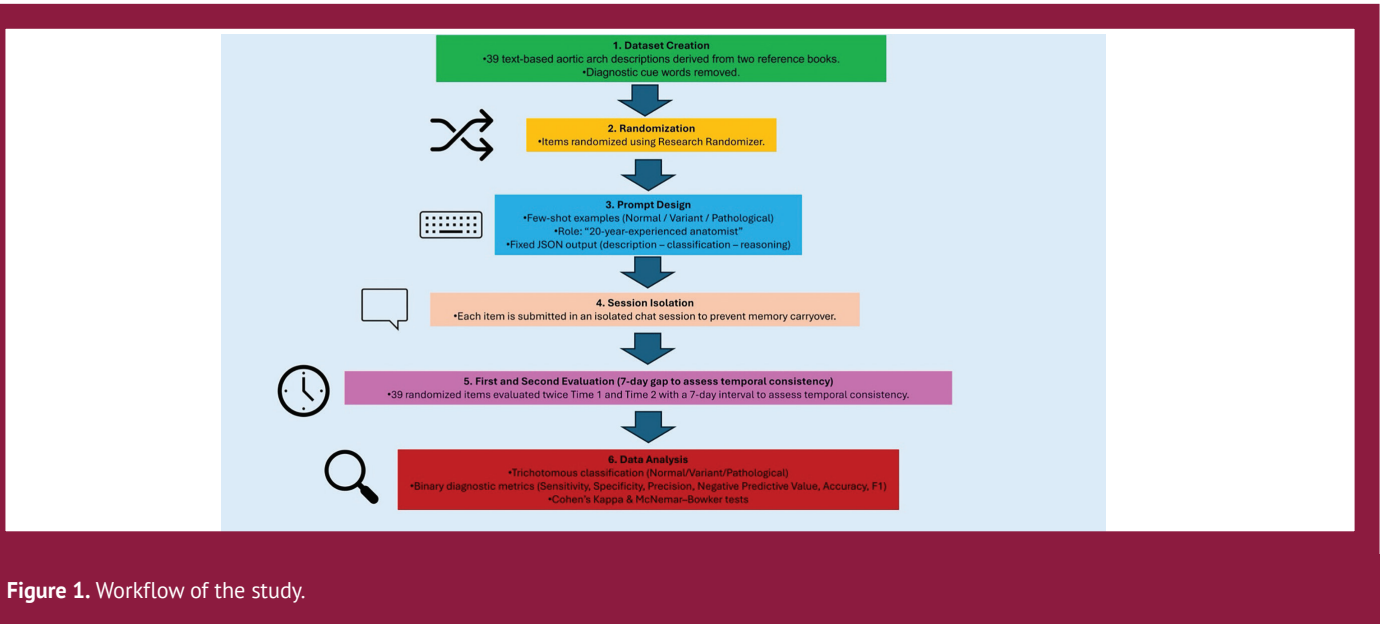


Figure 1. Workflow of the study.

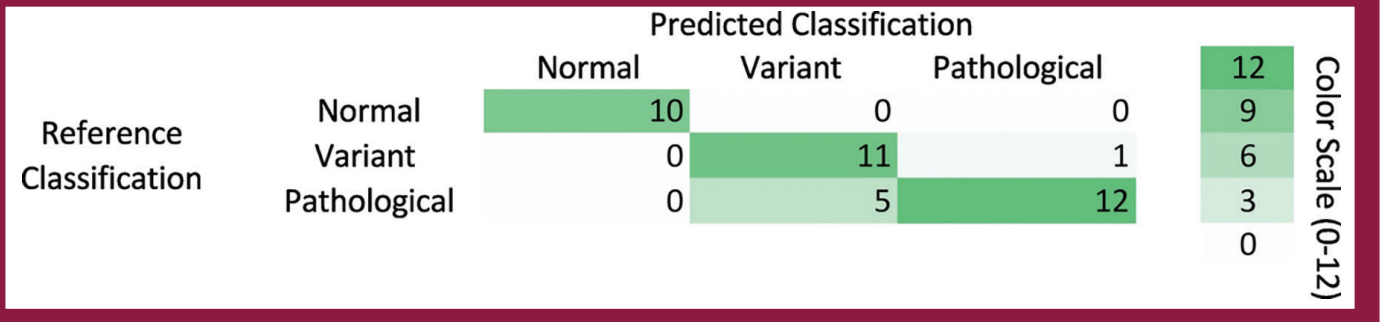


Figure 2. Heatmap visualization of model predictions compared with reference classifications.

Interrater agreement analysis demonstrated substantial overall concordance between ChatGPT and the reference books. The agreement was almost perfect for sentences describing normal anatomy. Substantial agreement was also present for pathological and variant sentences (Table 2).

The comparison of ChatGPT's classifications at the first and second evaluations showed no statistically significant differences. The McNemar–Bowker test indicated that the distribution of classifications remained stable across the two time points ($\chi^2 = 0.333$, $p = 0.564$).

In the binary diagnostic evaluation, sentences representing normal anatomy and anatomical variants were classified as non-pathologic, while sentences describing pathological findings were classified as pathologic (Table 3). These two categories were used to compute overall classification accuracy metrics, including sensitivity, specificity, precision, NPV, accuracy, and F1 score (Table 4). The model correctly classified 12 of 17 pathological descriptions and 21 of 22 non-pathological descriptions, corresponding to a sensitivity of 70.59%, a specificity of 95.45%, and an overall accuracy of 84.62%. The precision and NPV were 92.31% and 80.77%, respectively, while the F1 score was 80.00%. Within this dataset, the model showed higher specificity and precision than sensitivity, reflecting its tendency to classify some pathological descriptions as non-pathological. Given the limited number of items, particularly in the pathological category, these performance estimates should be interpreted cautiously.

Temporal consistency and item-level agreement were evaluated across the 39 items administered in two distinct

Table 1. Confusion matrix table showing the distribution of model predictions across the three diagnostic categories (row percentages shown).

Reference/responses from ChatGPT	Variant	Pathologic	Normal	Accuracy	p
Variant	11 (91.7%)	1 (8.3%)	0 (0%)	91.7%	0.102
Pathologic	5 (29.4%)	12 (70.6%)	0 (0%)	70.6%	
Normal	0 (0%)	0 (0%)	10 (100%)	100%	

*Test: McNemar Bowker test.

Table 2. Interrater agreement between ChatGPT and reference answers for each category and overall assessment.

Category	Conditional probability	Kappa (κ)	95% CI (lower–upper)	p value
Normal	1	1	0.692 - 1.0000[a]	< 0.001
Variant	0.786	0.666	0.406 – 0.885[b]	< 0.001
Pathological	0.8	0.675	0.414 – 0.889[b]	< 0.001
Overall agreement	–	0.766	0.571 – 0.921[b]	< 0.001

*Test: Cohen's Kappa. [a], Clopper-Pearson Exact Binomial; [b], Percentile Bootstrap Method (2000 resamples).

Table 3. Binary confusion matrix of pathological vs. non-pathological classifications.

Source\predicted	Pathologic	Non-pathologic (normal and variant)
Pathologic	12 (70.6%)	5 (29.4%)
Non-pathologic (normal and variant)	1 (4.5%)	21 (95.5%)

Table 4. Classification accuracy metrics of the ChatGPT-5.1 Plus classification compared with the reference standard.

Measure	Value (%)	Confidence interval (95%)
Sensitivity (recall)	70.59	44.04% – 89.69%*
Specificity	95.45	77.16% – 99.88%*
Precision (PPV)	92.31	63.97% – 99.81%*
Negative predictive value	80.77	60.65% – 93.45%*
Accuracy	84.62	69.47% – 94.14%*
F1 score	80	62.07% – 90.91%**

*Calculated using the Clopper-Pearson exact method for binomial proportions.
** Calculated using the non-parametric percentile bootstrap method.
PPV, positive predictive value.

sessions (November 15 and November 22, 2025). The overall observed percent agreement between the two sessions was 94.9% (37/39 items). Cohen's Kappa demonstrated almost perfect test-retest reliability (kappa = 0.922, p < 0.001), confirming strong item-level reproducibility over time (Table 5).

Across the 39 evaluated items, the model misclassified 6 cases. Of these errors, one configuration that should have been categorized as a benign variant was incorrectly labeled as pathological, while the remaining five pathological patterns were misidentified as simple variants. These errors represent a misclassification rate of 15.4% (6/39) overall, with 2.6% (1/39) reflecting overcalling of pathological abnormalities and 12.8% (5/39) reflecting undercalling of pathological abnormalities. In Table 6, each misclassified item is presented with its question description, accompanying reasoning, ChatGPT's classification, and the corresponding reference answer from the source material. The distribution and types of misclassifications are also presented in Figure 3.

Discussion

LLMs are increasingly used across various educational and clinical communication tasks (9). In recent years, studies incorporating image-based tasks have also begun to emerge (10,11). Furthermore, studies across various fields demonstrate the use of LLMs for clinical diagnostic purposes (4,12). In this context, our study examined how effectively these models recognize aortic arch variations, which are both anatomically and surgically important. We also assessed their ability to distinguish normal from pathological configurations by comparing their outputs with established reference sources. These findings should be interpreted in the context of the study's text-based design. The model classified isolated written descriptions rather than directly evaluating computed tomography angiography or other imaging data. Therefore, the study assessed semantic interpretation and categorical reasoning; the results cannot be directly generalized to radiological diagnosis or clinical image assessment.

Table 5. Item-level test–retest agreement and classification consistency between evaluation sessions.

1st Session	2nd Session				Overall agreement (%)	Cohen's kappa (κ)	p value
	Normal	Pathologic	Variant	Total			
Normal	10 (25.6%)	0 (0.0%)	1 (2.6%)	11 (28.2%)			
Pathologic	0 (0.0%)	12 (30.8%)	0 (0.0%)	12 (30.8%)	94.9% (37/39)	0.922	< 0.001
Variant	0 (0.0%)	1 (2.6%)	15 (38.5%)	16 (41.0%)			
Total	10 (25.6%)	13 (33.3%)	16 (41.0%)	39 (100.0%)			

Session 1 was conducted on November 15, 2025; Session 2 was conducted on November 22, 2025. Test: Cohen's Kappa.

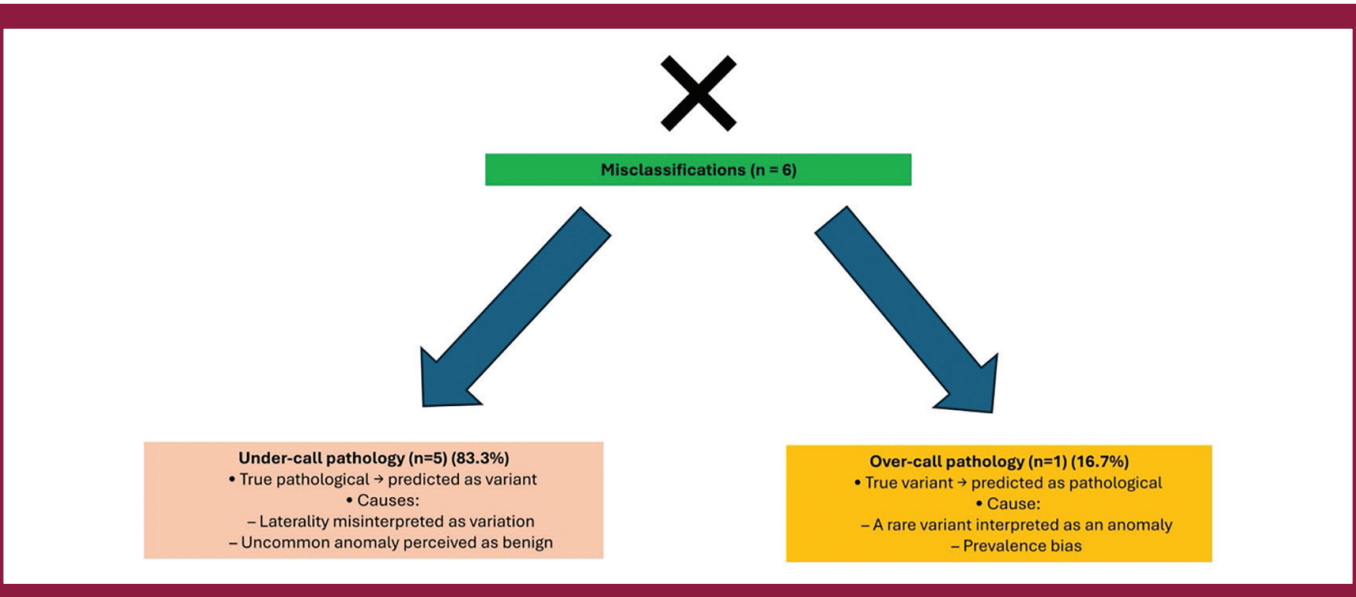


Figure 3. Distribution and types of misclassifications.

Prompt design also plays a critical role in the quality of the output. Advanced prompting strategies are often required to reduce hallucinations and maintain factual accuracy. Techniques such as one-shot or few-shot examples, assigning explicit roles, and defining a target audience can substantially influence model performance and improve consistency (13,14). Using a JSON-based output format helps constrain responses to a predefined structure, reducing variability and facilitating reliable analysis (15). In our study, we incorporated few-shot prompting, role-based prompting, and a JSON-based structured output template to enhance consistency and ensure standardized analysis. Because the same prompt was consistently applied across all items and sessions, the results reflect the model's classification

Table 6. Descriptions, reasoning, and correct classifications for items misclassified by ChatGPT.

Question number	Description	Reasoning	ChatGPT response (P[a]/V[b]/N[c])	Reference classification based on the source book[d]
12	The brachiocephalic artery arises from the descending aorta.	The brachiocephalic (innominate) artery normally arises as the first major branch of the aortic arch. Origin from the descending aorta is not a recognized anatomical variant and would represent a severe malposition inconsistent with normal embryological development of the aortic arch derivatives, indicating a pathological anomaly.	**P**	**V**
15	The aortic arch crosses over the right main bronchus due to the regression of the left dorsal aorta, leading to the descending aorta opacity along the right paraspinal space.	Regression of the left dorsal aorta with persistence of the right dorsal aorta produces a right-sided aortic arch that passes over the right main bronchus and descends on the right side of the vertebral column. This represents a recognized congenital laterality variant of the great vessels rather than the standard left-sided aortic arch, but it does not inherently constitute a pathological malformation in isolation.	**V**	**P**
17	A branching pattern where the left brachiocephalic artery is the first branch, followed by the right common carotid and the subclavian arteries.	A left brachiocephalic (innominate) artery arising as the first branch of the aortic arch represents a recognized but uncommon arch branching variant resulting from altered partitioning of the embryonic aortic sac and fourth aortic arch derivatives. Although unusual compared with the standard three-branch pattern, it is typically non-pathological.	**V**	**P**
23	An uncommon, congenital formation characterized by a kinking of a tortuous aortic arch at the level of the ligamentum arteriosum, which lacks the significant pressure gradient seen in aortic coarctation.	This configuration represents a congenital anatomic variant of the aortic arch in which tortuosity and kinking occur near the ligamentum arteriosum but without the hemodynamic obstruction characteristic of pathological coarctation. Because it does not produce the functional impairment or pressure gradient associated with true coarctation, it is best categorized as a structural vascular variant rather than a pathological anomaly.	**V**	**P**
25	The aortic arch is elongated, running upward, typically on the right side, above the sternum.	An aortic arch that ascends unusually high or courses on the right side represents a congenital deviation from the typical left-sided arch derived from the left fourth pharyngeal arch artery. A right-sided or high-riding aortic arch is a recognized vascular variant that may be asymptomatic, though it can coexist with other anomalies. In isolation, it is considered a variant rather than a pathological malformation.	**V**	**P**
27	A focal enlargement with smooth margins located along the anteroinferior surface of the isthmic region of the aortic arch, often considered a vestige of the closed ductus arteriosus.	This description matches a ductus diverticulum, a smooth-margined outpouching at the aortic isthmus representing a benign remnant of the closed ductus arteriosus. It is a recognized anatomical variant and not indicative of pathological aneurysmal change.	**V**	**P**

[a], pathologic; [b], variant; [c], normal; [d], Bergman's comprehensive encyclopedia of human anatomic variation (7).

performance under controlled and reproducible prompting conditions. Nevertheless, performance may differ with alternative prompting strategies.

Recent research demonstrates that LLMs achieve variable, but often promising, performance on anatomy-related tasks. In a musculoskeletal radiology study involving 175 questions from Radiopaedia, ChatGPT-4o achieved an overall accuracy of 69.7% and performed exceptionally well on text-based questions (88.2%), even surpassing radiologists in specific anatomical regions. However, it consistently underperformed in image-based interpretation (10). Similarly, in our study, ChatGPT-5.1 Plus achieved an accuracy of 84.62% on all text-based questions.

Totlis et al. (16) reported that ChatGPT-4 provided generally accurate and well-structured anatomical explanations; however, its performance declined when addressing anatomical variants unless these were explicitly categorized into specific types. The authors noted that the model often omitted variant-related details and produced less reliable responses when variation patterns were complex (16). These observations align with our findings, as the model frequently misinterprets pathological patterns as anatomical variants. Nevertheless, it accurately identified aortic arch variations in 11 out of 12 cases (91.7%). Expanding similar investigations across different topographic regions will be essential for strengthening the literature on anatomical variation and for understanding the limits of LLM-based anatomical classification.

The misclassifications observed in this study suggest that the model frequently interpreted descriptors, such as an anatomical structure being "on the right" or "higher than expected," and clinical descriptors such as "hemodynamically stable," as indicators of benign variation rather than pathology. As a result, configurations that deviated from expected laterality or course were incorrectly labeled as variants. Conversely, a genuinely rare variant, such as a brachiocephalic artery arising from the descending aorta, was classified as pathological because of its low prevalence. However, some disagreements may also reflect differences in anatomical terminology and classification among sources. Therefore, the reference categories used in this study were defined based on the selected source materials. Overall, despite achieving perfect accuracy on normal configurations, the model had greater difficulty distinguishing uncommon variants from pathological anomalies.

A review study analyzing 53 studies found that ChatGPT-4 achieves significantly higher accuracy than ChatGPT-3.5 on medical licensing and residency examinations (17). In another study, a comparative analysis of eight contemporary LLMs for dental specialty examinations demonstrated that ChatGPT-o1 outperformed all other models, including Gemini 1.5, which showed the lowest accuracy (18). In line with recent comparative findings, Gemini 2.5 Flash demonstrated the strongest performance on anatomy-focused multiple-choice questions, outperforming other tested models, including ChatGPT-4o, and DeepSeek V3 (19). These findings underscore that performance varies considerably across model generations, suggesting that ongoing updates to and advancements in LLM architectures may yield different or improved outcomes over time.

In a study, ChatGPT-4o and ChatGPT-4o mini were evaluated for imaging analysis of 53 lumbar cases, comprising both patients and healthy individuals. ChatGPT-4o achieved an accuracy of 0.81, with precision, recall, and F1 scores all exceeding 0.80, and a Kappa value of 0.61, reflecting moderate agreement with orthopedic surgeons (20). When analyzing 117 knee radiographs from patients with osteoarthritis, ChatGPT-4o correctly identified all images and demonstrated strong binary classification performance, achieving a precision of 0.95, recall of 0.83, and an F1 score of 0.88. However, its detailed grading accuracy remained markedly lower than that of orthopedic surgeons (21). These findings parallel our results, which showed that ChatGPT maintained high sensitivity and specificity for common patterns but exhibited a marked decline in accuracy when confronted with rare vascular variations or anatomically subtle distinctions.

The model demonstrated an excellent ability to recognize normal and variant anatomy, but showed comparatively lower performance in detecting pathological configurations. In the binary diagnostic evaluation, this pattern translated into high specificity (95.45%) and precision (92.31%), indicating a strong ability to identify non-pathological cases. In contrast, the sensitivity (70.59%) reflects the model's tendency to miss a proportion of pathological cases.

Study Limitations

This study has several significant limitations. Although detailed, carefully structured prompts were used, all questions were presented solely in text, without accompanying imaging data. Therefore, the findings do not provide insight into the model's image-based classification accuracy. Furthermore, the dataset comprised 39 descriptions, reflecting the number of eligible aortic arch configurations identified in the selected reference sources within the predefined scope of the study. Nevertheless, the limited sample size may reduce the precision and generalizability of the findings. As LLMs are frequently updated, the current findings are specific to the ChatGPT-5.1 Plus version used during the study period and may not generalize to future iterations. Another limitation is that the study assessed discrete, isolated anatomical descriptions rather than full clinical scenarios. Therefore, the model's performance

in integrating multimodal clinical information, such as symptoms or hemodynamic parameters, was not tested.

Conclusion

These findings are limited to text-based classification and should not be interpreted as evidence of the model's ability to analyze CT angiography or perform image-based clinical diagnosis. ChatGPT-5.1 Plus accurately identified normal aortic arch anatomy and the more common variants. However, its sensitivity declined when evaluating rare variants and pathological configurations. These results indicate that the model has substantial potential to support anatomical interpretation. Furthermore, its limitations, especially in recognizing clinically meaningful anomalies, remain evident. Therefore, expert oversight is still essential when using such systems in educational or diagnostic settings.

Ethics

Ethics Committee Approval: This article does not contain any studies with human participants or animals. Ethical approval was not required for this type of publication.

Informed Consent: Informed consent was not required because the article does not contain any studies with human participants or animals.

Footnotes

Authorship Contributions

Concept: İ.S., T.Ç., Design: İ.S., F.M.Ö., T.Ç., Data Collection or Processing: İ.S., F.M.Ö., G.G., E.K., Analysis or Interpretation: İ.S., F.M.Ö., G.G., E.K., Literature Search: İ.S., Writing: İ.S., F.M.Ö., G.G., E.K.

Conflict of Interest: No conflict of interest was declared by the authors.

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

SupplementaryFileLinks:https://d2v96fxpocvxx.cloudfront.net/f942abf9-a73b-4057-b48b-c3a4b7d0c2a5/content-images/b7aa1347-2a23-45a0-8528-3bd783eab98d.pdf

REFERENCES

1. Bae SB, Kang EJ, Choo KS, Lee J, Kim SH, Lim KJ, et al. Aortic arch variants and anomalies: embryology, imaging findings, and clinical considerations. J Cardiovasc Imaging. 2022;30:231-262. [Crossref]
2. Xaviour R, Joseph KK, Jacob JT. Anatomical variations and embryological basis of arch of aorta and aortic valve. Anat Sci Int. 2024;99:305-319. [Crossref].
3. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. JMIR Med Inform. 2024;12:e56426. [Crossref]
4. Mazzucchelli M, Salzano S, Caltabiano R, Magro G, Certo F, Barbagallo G, et al. Diagnostic performance of ChatGPT-4.0 in histopathological analysis of gliomas: a single institution experience. Neuropathology. 2025;45:e70023. [Crossref]
5. Sozer A, Sahin MC, Sozer B, Erol G, Tufek OY, Nernekli K, et al. Do LLMs have 'the Eye' for MRI? Evaluating GPT-4o, Grok, and Gemini on brain MRI performance: first evaluation of Grok in medical imaging and a comparative analysis. Diagnostics (Basel). 2025;15:1320. [Crossref]
6. Standring S, editor. Gray's anatomy: the anatomical basis of clinical practice. 41st ed. London: Elsevier; 2016. [Crossref]
7. Tubbs RS, Shoja MM, Loukas M, editors. Bergman's comprehensive encyclopedia of human anatomic variation. Hoboken (NJ): John Wiley & Sons; 2016. p. 501-529. [Crossref]
8. Urbaniak GC, Plous S. Research randomizer (version 4.0) [Internet]. [cited 2025 Nov 12]. [Crossref]
9. Temizsoy Korkmaz F, Ok F, Karip B, Keleş P. Exploring body donation communication with large language models: Accuracy, readability, and ethical considerations. Anat Sci Educ. 2025;18:1238-1249. [Crossref]
10. Salbas A, Baysan EK. Assessment of large language models in musculoskeletal radiological anatomy: a comparative study with radiologists. Jt Dis Relat Surg. 2026;37:190-199. [Crossref]
11. Sarangi PK, Datta S, Panda BB, Panda S, Mondal H. Evaluating ChatGPT-4's performance in identifying radiological anatomy in FRCR part 1 examination questions. Indian J Radiol Imaging. 2024;35:287-294. [Crossref]
12. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology. 2023;308:e231040. [Crossref]
13. Kıyak YS. Beginner-level tips for medical educators: guidance on selection, prompt engineering, and the use of artificial intelligence chatbots. Med Sci Educ. 2024;34:1571-1576. [Crossref]
14. Shah K, Xu AY, Sharma Y, Daher M, McDonald C, Diebo BG, et al. Large Language model prompting techniques for advancement in clinical medicine. J Clin Med. 2024;13:5101. [Crossref]
15. Sayeed HM, Mohanty T, Sparks TD. Annotating materials science text: a semi-automated approach for crafting outputs with Gemini Pro. Integr Mater Manuf Innov. 2024;13:445-452. [Crossref]
16. Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 2023;45:1321-1329. [Crossref]
17. Jaleel A, Aziz U, Farid G, Zahid Bashir M, Mirza TR, Khizar Abbas SM, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. JMIR Med Educ. 2025;11:e68070. [Crossref]
18. Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE. Evaluation of the performance of different large language models on head and neck anatomy questions in the dentistry specialization exam in Turkey. Surg Radiol Anat. 2025;47:211. [Crossref]
19. Singal A, Goyal S. Comparative evaluation of AI platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o" in solving multiple-choice questions from different subtopics of anatomy. Surg Radiol Anat. 2025;47:193. [Crossref]
20. Wang S, Wang Y, Jiang L, Chang Y, Zhang S, Zhao K, et al. Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. Eur J Med Res. 2025;30:45. [Crossref]
21. Zhu J, Jiang Y, Chen D, Lu Y, Huang Y, Lin Y, et al. High identification and positive-negative discrimination but limited detailed grading accuracy of ChatGPT-4o in knee osteoarthritis radiographs. Knee Surg Sports Traumatol Arthrosc. 2025;33:1911-1919. [Crossref]