applied sciences

MDPI

Article

Comparing ChatGPT Assessments Based on Transcripts and Screenshots with Anesthesiologists' Evaluations of Ultrasound-Guided Abdominal Wall Block Videos: A Cross-Sectional Study

Nuran Akıncı Ekinci * and Esma Karaarslan

Department of Anesthesiology and Reanimation, Konya City Hospital, Konya 42020, Turkey
* Correspondence: nurandr@yahoo.com

Highlights

Public health relevance: How does this work relate to a public health issue?

Online video platforms are increasingly used for medical education, yet the variable quality of procedural content may affect the reliability of information available to healthcare professionals.

Scalable approaches to evaluating online educational content may help identify potentially inadequate or unreliable resources.

Public health significance: Why is this work of significance to public health?

This study provides empirical evidence on how closely a large language model aligns with experienced anesthesiologists when evaluating the quality and reliability of ultrasound-guided regional anesthesia videos.

Although total scores did not differ significantly, agreement with experts was limited, and item-level score differences were observed for specific technical and safety-related aspects, highlighting important limitations of automated evaluation.

Public health implications: What are the key implications or messages for practitioners, policy makers and/or researchers in public health?

Abstract

Digital video-sharing platforms are widely used in anesthesia education, but the educational quality of ultrasound-guided abdominal wall block (UAWB) videos is heterogeneous. This cross-sectional study assessed agreement between a large language model (LLM) and experienced anesthesiologists in evaluating UAWB videos on YouTube. Seventy-five videos covering five UAWB techniques were independently evaluated by two anesthesiologists and ChatGPT version 5.2 using a researcher-developed 10-item Educational Content Checklist (ECC), DISCERN, the Global Quality Scale (GQS), and JAMA benchmark criteria. The LLM received each video's full transcript, metadata screenshot, and 5–9 procedural screenshots rather than the full video. Human–LLM scores were compared using the Wilcoxon signed-rank test, and agreement was quantified using intraclass correlation coefficients. Total scores did not differ significantly for the ECC (median 34 vs. 32; $p = 0.168$), DISCERN (54 vs. 54; $p = 0.347$), GQS (4 vs. 4; $p = 0.469$), or JAMA criteria (2 vs. 2; $p = 0.865$). However, agreement was limited: moderate for ECC (ICC = 0.576, 95% CI 0.404–0.709) and JAMA (ICC = 0.529, 95% CI 0.343–0.674), and poor for GQS (ICC = 0.495, 95% CI 0.303–0.648) and DISCERN (ICC = 0.357, 95% CI 0.146–0.538). In exploratory item-level analyses, differences emerged in vascular safety, needle technique, and indications. The absence of significant

Check for updates

Academic Editor: Jie Yang

Received: 31 August 2026
Revised: 15 October 2026
Accepted: 27 October 2026
Published: 28 October 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) license.

differences in total scores should not be interpreted as equivalence. LLM-assisted prescreening is a potential application that requires prospective evaluation, and the present findings do not support replacing expert assessment with LLM evaluation, particularly for dynamic, safety-critical technical content.

Keywords: regional anesthesia; video recording; artificial intelligence; large language models; education; medical; quality assurance; health care; Youtube

1. Introduction

Large language models may have the potential to assist with the initial screening of large volumes of online medical educational content; this requires prospective validation, and LLMs should not replace expert assessment.

Expert oversight remains particularly important when evaluating dynamic, safety-critical procedural content, and should be retained in any future AI-assisted quality assessment approach.

Digital educational resources, particularly video-sharing platforms such as YouTube, have become an integral component of medical education for residents and practicing physicians. This is particularly evident in technically demanding disciplines such as regional anesthesia, where trainees frequently rely on procedural videos [1,2].

Ultrasound-guided abdominal wall blocks—such as transversus abdominis plane (TAP), rectus sheath (RSB), erector spinae plane (ESP), quadratus lumborum (QL), and ilioinguinal–iliohypogastric (II–IH)—are important components of multimodal analgesia; for example, TAP block has been shown to reduce postoperative opioid use [3]. However, safe and effective performance of these techniques requires advanced proficiency in sonoanatomy and needle handling [4]. Consequently, ensuring the quality and accuracy of educational videos for these procedures is clinically important.

Previous studies have demonstrated considerable variability in the quality and reliability of medical educational videos available on YouTube, often providing incomplete or inconsistent information [5–7]. Several validated instruments, including DISCERN, the Global Quality Scale (GQS), and the JAMA benchmark criteria, are commonly used to evaluate educational quality and information reliability [8]. However, these instruments require expert manual assessment, which is time-consuming, resource-intensive, and subject to inter-rater variability.

Recent advances in large language models (LLMs) have created new opportunities for the automated evaluation of medical information and educational resources. However, evidence regarding the ability of LLMs to provide evaluations comparable to those of clinical experts remains limited, particularly for procedural educational videos in regional anesthesia [9]. Previous studies comparing ChatGPT with human experts in evaluating YouTube videos have reported varying results depending on the assessment instruments and clinical context. For tonsillectomy videos, accuracy and completeness scores generated by ChatGPT-4 from transcripts correlated significantly with expert JAMA and DISCERN scores [9]. For fall-prevention exercise videos for older adults, GQS scores did not differ significantly, whereas ChatGPT-4.5 assigned significantly higher DISCERN scores than the experts [10]. For urological cancer videos, no significant differences were found in DISCERN-5 and GQS scores, although the repeatability of ChatGPT evaluations was low [11]. A more recent study of anesthesia education videos reported strong correlations between ChatGPT-5 evaluations based on transcripts and metadata and expert scores [12]. However, correlation alone does not demonstrate agreement between methods. To our knowledge, no previous study has systematically compared LLM-based and expert evaluations of ultrasound-guided

abdominal wall block educational videos using multiple quality assessment instruments, including a procedure-specific checklist covering safety-critical steps, while providing the model with selected visual inputs in addition to transcripts. Therefore, this study aimed to determine whether an LLM (ChatGPT, version 5.2) evaluates the educational quality and information reliability of ultrasound-guided abdominal wall block videos on YouTube differently from experienced anesthesiologists, and to quantify the level of agreement between the two methods. The evaluation was conducted from the perspective of anesthesia trainees and clinicians performing these procedures.

We hypothesized that ChatGPT would demonstrate no significant difference in overall evaluation scores but lower agreement with experts in procedure-specific technical domains.

2. Materials and Methods

2.1. Ethics Approval and Reporting Standards

This descriptive cross-sectional study analyzed publicly available YouTube videos demonstrating ultrasound-guided abdominal wall blocks. The study was reported in accordance with the applicable items of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement for cross-sectional studies. Because the study involved only publicly accessible online content and associated metadata and did not include human participants, animals, identifiable personal data, or clinical interventions, institutional ethics committee approval and clinical trial registration were not required.

2.2. Video Search Strategy and Selection Process

A systematic search of YouTube was conducted on 23 December 2025, and eligible videos were identified and their URLs and characteristics were recorded on the same date to minimize variability arising from changes in YouTube content or search rankings over time. All expert and LLM evaluations were performed between 24 December 2025 and 6 January 2026. To reduce the effects of personalized algorithms, all searches were performed in incognito mode, with browser history and cookies cleared before each search. Five ultrasound-guided abdominal wall blocks were evaluated: transversus abdominis plane (TAP), rectus sheath (RSB), erector spinae plane (ESP), quadratus lumborum (QL), and ilioinguinal–iliohypogastric (II-IH) nerve blocks. Each block type was searched separately using the following predefined keywords: "transversus abdominis plane block ultrasonography," "rectus sheath block ultrasonography," "erector spinae plane block ultrasonography," "quadratus lumborum block ultrasonography," and "ilioinguinal iliohypogastric block ultrasonography." Search results were sorted in descending order according to view count. Videos were screened according to predefined inclusion and exclusion criteria, and the selection process is illustrated in Figure 1. No formal sample size calculation was performed; for each of the five UAWB techniques, the 15 most-viewed videos meeting the inclusion criteria were selected, yielding a total of 75 videos. View count was used as a sampling criterion reflecting the visibility of videos on the platform rather than as an indicator of quality; a similar purposive sampling approach has been used in previous content analyses of YouTube videos [10,11]. To ensure that each of the five block techniques was equally represented in the study sample, 15 videos were selected per technique. As no formal sample size calculation was performed, the study should be regarded as exploratory, and the sample is not intended to represent all educational videos on YouTube.

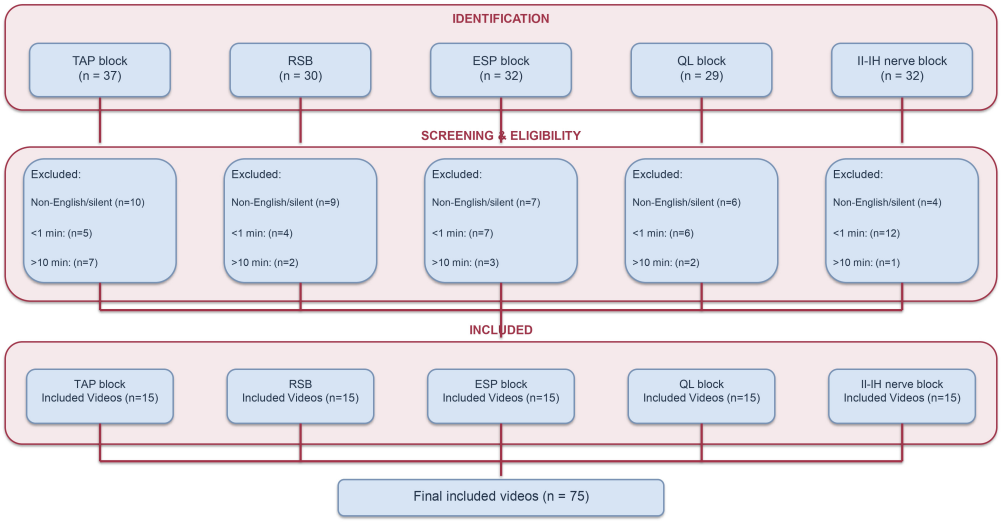


Figure 1. Flow diagram of the study. Flow diagram illustrating the identification, screening, eligibility assessment, and final inclusion of YouTube videos related to ultrasound-guided abdominal wall blocks.

2.3. Inclusion and Exclusion Criteria

Videos were included if they provided educational or procedural content related to ultrasound-guided TAP, RSB, ESP, QL, or II–IH blocks; had a duration between 1 and 10 min; contained English voice narration and a verifiable English transcript; and were published on YouTube as original, full-length content. English-language videos were included because English is widely used in international medical education and these videos are therefore accessible to a broad international audience of trainees and clinicians. The requirement for an available transcript ensured that the complete transcript of each video could be used in the LLM evaluation. The 1–10 min duration limit was applied to keep the evaluated videos within a comparable length range and to facilitate their representation to the LLM by a limited number of screenshots. Videos shorter than 1 min or longer than 10 min, silent videos or those consisting of visuals only, videos without available transcripts, content produced in languages other than English, videos unrelated to the procedure or showing discordance between the title and content, as well as duplicate videos or re-uploads of the same content, were excluded from the study.

2.4. Video Characteristics and Quantitative Variables

For each included video, several quantitative variables reflecting general video characteristics and viewer engagement were recorded. These variables included video duration (minutes), the number of subscribers of the uploading channel, upload time, total view count, number of likes, and number of comments. Channel subscriber counts were obtained from the YouTube homepage of the respective channels.

Upload time was calculated in months based on the interval between the date the video was uploaded to YouTube and the date of data collection.

Viewer engagement was quantified using the engagement rate, which was calculated for each video using the following formula:

$$\text{Engagement rate}\ (\%) = [(\text{Number of likes} + \text{Number of comments})/\text{Number of views}] \times 100$$

To assess viewing performance over time, the view rate was calculated by dividing the total number of views by the number of days elapsed since the video was uploaded:

$$\text{View rate}\ (\text{views/day}) = \text{Total number of views}/\text{Number of days since upload}$$

These quantitative variables were analyzed to evaluate overall video popularity, visibility, and viewer engagement levels.

2.5. Human Expert Evaluation

All videos included in the study were independently evaluated by two anesthesiologists experienced in ultrasound-guided regional anesthesia. The two evaluators were the authors (N.A.E. and E.K.). The evaluators were blinded to each other's scores and to the results of the LLM evaluation. To minimize assessment bias, all evaluations were performed independently using predefined scoring criteria, and the LLM was provided with a standardized set of inputs for every video (detailed below). Videos were viewed in their original format on the YouTube platform using the recorded URLs. Each evaluator performed the assessments independently using the predefined scoring criteria. Inter-rater agreement was quantified using ICC (two-way mixed-effects, absolute agreement, single measures) and ranged from good to excellent across all evaluation scales [13]. ICC = 0.837 (95% CI 0.718–0.903) for the 10-item structured Educational Content Checklist, ICC = 0.887 (95% CI 0.827–0.928) for the total DISCERN score, ICC = 0.830 (95% CI 0.744–0.889) for GQS, and ICC = 0.904 (95% CI 0.852–0.938) for the JAMA benchmark criteria (all $p < 0.001$).

2.6. LLM-Based Evaluation

After completion of the human expert evaluations, the same 75 videos were independently assessed using a large language model (ChatGPT, version 5.2; OpenAI, San Francisco, CA, USA). The model was accessed via the web interface (https://chatgpt.com, Plus subscription) within the study evaluation period (24 December 2025 to 6 January 2026); the API was not used. Default model settings were used, as parameters such as temperature cannot be adjusted in this interface. The model was provided with a standardized evaluation prompt defining the scoring instruments and assessment criteria full prompt reproduced verbatim in Supplementary File S1. The same prompt was used for all evaluations. For each video, the model received the complete transcript, one metadata screenshot (video title, channel name, subscriber count, view count, number of likes, and upload date), and 5–9 procedural screenshots representing key stages of the procedure. Transcripts were obtained from YouTube's transcript function and used without manual correction to avoid researcher influence on the LLM input and to reflect the input available in a real-world automated workflow. Screenshots were selected after completion of the expert evaluations by an anesthesiologist who was not involved in the expert scoring, according to predefined categories: patient and operator positioning, ultrasound probe placement, sonoanatomical visualization of the target plane and adjacent structures, Doppler imaging of vascular structures when applicable, needle approach and visualization of the needle tip, and local anesthetic spread (hydrodissection). Screenshot timestamps were not recorded. The same anesthesiologist submitted the transcripts and screenshots to the model. Evaluations were based on these combined textual and visual inputs, and the model was instructed to rely exclusively on the information explicitly provided, without inferring or assuming missing details. A new conversation was initiated for each video to prevent contextual carryover between evaluations, and no iterative prompting, follow-up instructions, or score revision was performed after the initial response. Each video was evaluated once, and repeated runs were not performed. The LLM therefore assessed a multimodal representation of each video (transcript, metadata, and selected screenshots) rather than the continuous video itself.

2.7. Evaluation Scales and Scoring

Both human experts and the LLM independently evaluated all videos using the same four assessment tools: the Educational Content Checklist (ECC), DISCERN, the Global Quality Scale (GQS), and the JAMA benchmark criteria. Given that the evaluation was conducted from the perspective of anesthesia trainees and clinicians performing these procedures, the instruments were chosen to cover complementary aspects: procedure-specific educational content (ECC), information reliability and the quality of treatment-related information (DISCERN), overall educational quality (GQS), and source transparency (JAMA).

Educational content quality was assessed using the Educational Content Checklist (ECC). The ECC was developed based on published guidelines and previous literature on ultrasound-guided regional anesthesia and refined through expert consensus to capture procedure-specific educational elements not adequately addressed by existing generic educational quality instruments. Its ten items cover indications and positioning, contraindications, sterility and preparation, sonoanatomy, vascular safety and Doppler use, probe handling and image optimization, needle technique, local anesthetic spread, pharmacology and dosing, and complication management. The checklist was developed by the authors and finalized before any scoring began, and it was not modified during the expert or LLM evaluations. Each item was scored on a 5-point Likert scale (1 = no information, 5 = very good).

Information reliability and the quality of treatment-related content were evaluated using the DISCERN instrument [14]. DISCERN is a 16-item instrument originally developed to assess written consumer health information on treatment choices; it was used because it is frequently applied in studies of YouTube medical videos, allowing comparison with previous work, although some of its patient-centered items are of limited relevance to professional procedural videos. Overall educational quality was assessed using the Global Quality Scale (GQS), scored on a 5-point Likert scale (1 = no information, 5 = very good) [15]. The GQS provides a single global rating of overall quality, flow, and usefulness. In addition, transparency and reliability indicators were evaluated using the JAMA benchmark criteria, including authorship, attribution, disclosure of conflicts of interest, and currency, each scored as present (1) or absent (0) [16].

Detailed item definitions and full scoring criteria for all assessment tools are available in Supplementary Files S2–S5. Item-level and total scores generated by the LLM were recorded separately for each video and compared with human expert evaluations.

2.8. Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics version 25.0 (IBM Corp., Armonk, NY, USA). Continuous variables were presented as median (25th–75th percentiles). The normality of data distribution was assessed using the Shapiro–Wilk test; distributional characteristics were further supported by examination of histograms and skewness and kurtosis values. As the variables were not normally distributed, non-parametric analyses were performed.

As the same YouTube videos were scored by both human experts and the LLM, comparisons were conducted using a paired data structure. For each video, the mean of the two experts' scores was used to represent the human evaluation. Given the lack of normal distribution, differences between human and LLM-based evaluations were analyzed using the Wilcoxon signed-rank test. For each comparison, the Hodges–Lehmann median difference and its 95% confidence interval were calculated as a measure of effect size. Analyses were performed separately for each evaluation scale, including the 10-item structured Educational Content Checklist, total DISCERN score, GQS, and JAMA benchmark criteria.

In addition, agreement between human expert and LLM-based evaluations (inter-method agreement) was assessed using ICC (two-way mixed-effects, absolute agreement,

single measures; ICC,(A,1)) with 95% confidence intervals for the total scores of each evaluation instrument. ICC values were interpreted according to Koo and Li as poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), or excellent (>0.90) [13]. Because the study was not designed as an equivalence study, a non-significant difference in total scores was not interpreted as evidence of equivalence; differences in scores and agreement between methods were treated as separate questions. Item-level comparisons were performed with the Wilcoxon signed-rank test and were considered exploratory; to account for multiple testing, p values were adjusted within each instrument using the Holm method [17]; adjusted p values are reported for nominally significant items. All statistical tests were two-sided, and a p value of <0.05 was considered statistically significant.

3. Results

The general characteristics of the 75 YouTube videos included in the study are summarized in Table 1. The upload time, view count, and quantitative variables reflecting viewer engagement showed wide variability across the videos. All quantitative variables were reported as median (25th–75th percentiles) and minimum–maximum values.

The results of the comparison between quality evaluations performed by human experts and the LLM for the same YouTube videos are presented in Table 2. No statistically significant difference was observed between human and LLM-based evaluations with respect to the total scores of the 10-item structured Educational Content Checklist ($p = 0.168$). Similarly, no statistically significant differences were found between the two evaluation modalities in terms of total DISCERN scores ($p = 0.347$), GQS scores ($p = 0.469$), or JAMA benchmark scores ($p = 0.865$).

Table 1. General characteristics of the videos.

Variable	Median (25th–75th Percentiles)	Minimum–Maximum
Subscriber count	8260 (1810–34,300)	3–250,000
Video duration (minutes)	4.13 (2.37–5.38)	1.10–10.00
Upload time (months)	55 (38–82)	1–180
Like count	122 (13–225)	0–3300
View count	14,094 (2211–43,569)	8–381,688
Comment count	3 (0–9)	0–63
Engagement rate (%)	0.85 (0.52–1.29)	0–10.39
View rate (views/day)	10.03 (1.18–24.96)	0.01–180.79

Data are presented as median (25th–75th percentiles) and minimum–maximum values.

Table 2. Comparison of quality assessment scores between human experts and the LLM.

	Human Evaluation	LLM		
Video Evaluation Parameters	**Human Experts**	**LLM**	**Paired Difference (95% CI)**	***p* Value**
Educational Content Score	34 (29–42)	32 (27–40)	−1.5 (−3 to 0.5)	0.168
DISCERN Score	54 (47–64)	54 (44–63)	−1 (−4.5 to 1.5)	0.347
Global Quality Score (GQS)	4 (3–5)	4 (3–4)	0 (−0.5 to 0)	0.469
JAMA Score	2 (2–3)	2 (2–3)	0 (0 to 0)	0.865

Data are presented as median (25th–75th percentiles). Paired differences are expressed as Hodges–Lehmann median differences with 95% confidence intervals. Comparisons between human expert and LLM evaluations were performed using the Wilcoxon signed-rank test. Statistical significance was set at $p < 0.05$. LLM, large language model; CI, confidence interval.

Item-level analysis of the Educational Content Checklist showed no statistically significant differences between human experts and the LLM for most domains (Figure 2). However, significant differences (unadjusted p values) were observed for three items. The LLM assigned higher scores for appropriateness of indications ($p = 0.001$), but lower scores for vascular safety and Doppler use ($p < 0.001$) and needle technique ($p = 0.003$). No significant differences were observed for contraindications, sterility and preparation, sonoanatomy, probe handling and image optimization, local anesthetic spread (hydrodissection), pharmacology and dosing, or complication management (all $p > 0.05$). After Holm correction, all three items remained statistically significant (adjusted $p = 0.009$, <0.001, and 0.021, respectively).

In the item-level analysis of the DISCERN instrument, no statistically significant differences between human experts and the LLM were detected for most items (Figure 3). However, nominally significant differences (before correction for multiple comparisons) were identified for several items, particularly within the second section assessing the quality of treatment-related information.

According to the results of the Wilcoxon signed-rank test (unadjusted p values), the LLM assigned significantly lower scores than human experts for the explanation of how the treatment works (Item 9; $p = 0.003$). Similarly, statistically significant differences were observed between the two evaluation modalities for the description of treatment benefits (Item 10; $p = 0.023$), explanation of what would happen if no treatment is provided (Item 12; $p = 0.033$), discussion of effects on quality of life (Item 13; $p < 0.001$), and support for shared decision-making (Item 15; $p = 0.047$). After Holm correction, only Item 13 remained statistically significant (adjusted $p = 0.007$); the difference for Item 9 was borderline (adjusted $p = 0.051$), and the differences for Items 10, 12, and 15 were not significant (adjusted $p = 0.319$, 0.426, and 0.564, respectively).

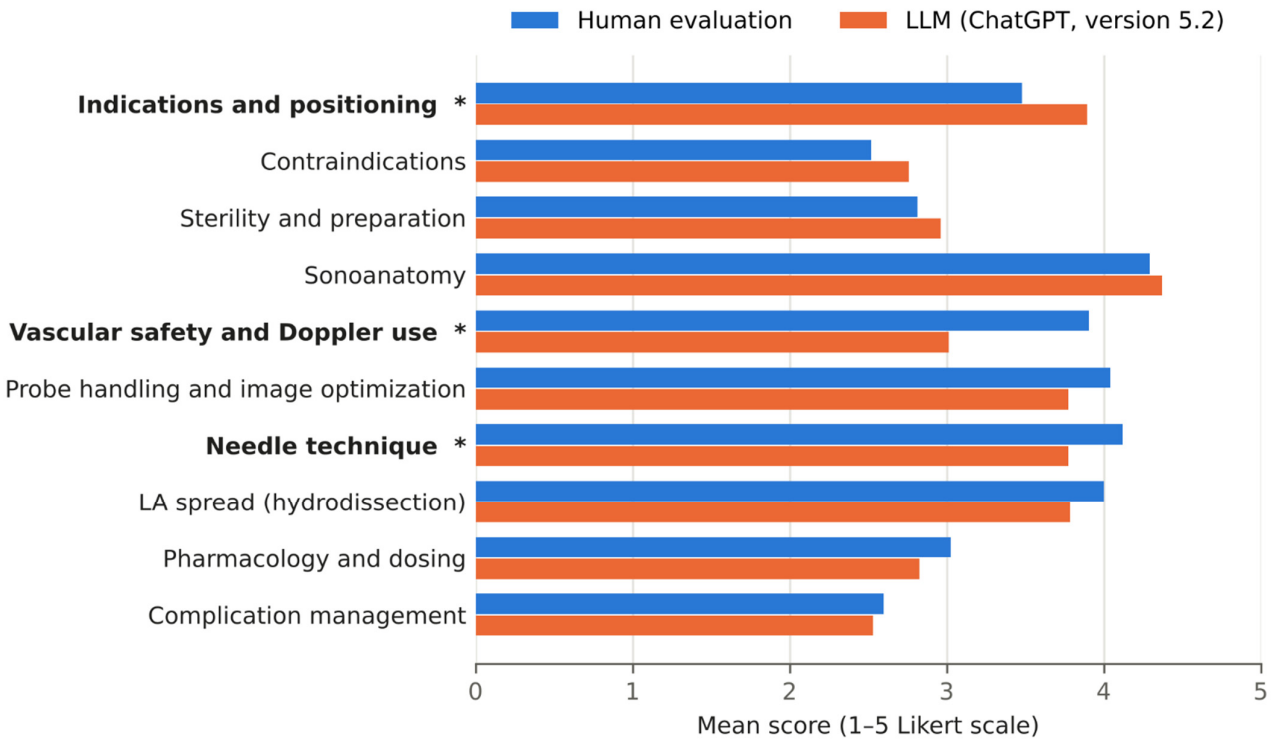


Figure 2. Comparison of item-level educational content scores. Comparison of mean item-level educational content scores between the human evaluation and the LLM (ChatGPT, version 5.2) using a 5-point Likert scale. LA: local anesthetic. Asterisks indicate items that remained significant after Holm correction for multiple comparisons (Wilcoxon signed-rank test).

In contrast, no statistically significant differences were observed for clarity of aims (Item 1), achievement of aims (Item 2), relevance to the topic (Item 3), citation of sources (Item 4), currency of information (Item 5; $p = 0.064$), impartiality (Item 6), provision of additional sources of information (Item 7), or discussion of areas of uncertainty (Item 8). Furthermore, no statistically significant differences were found between human and LLM-based evaluations regarding treatment risks (Item 11), alternative treatment options (Item 14), or the overall reliability score (Item 16) (all $p > 0.05$).

Comparisons of JAMA benchmark criterion scores between human expert and LLM evaluations are shown in Figure 4. No statistically significant differences were identified for any individual JAMA benchmark criterion (all $p > 0.05$).

Inter-method agreement between human experts and the LLM was evaluated using ICC analysis across all evaluation scales. For educational content scores, a moderate level of agreement was observed between human experts and the LLM, with an ICC value of 0.576 (95% CI 0.404–0.709; $p < 0.001$). Agreement was moderate for the JAMA benchmark criteria (ICC = 0.529, 95% CI 0.343–0.674; $p < 0.001$) and poor for the GQS (ICC = 0.495, 95% CI 0.303–0.648; $p < 0.001$). Agreement was also poor for total DISCERN scores assessing information reliability (ICC = 0.357, 95% CI 0.146–0.538; $p = 0.001$).

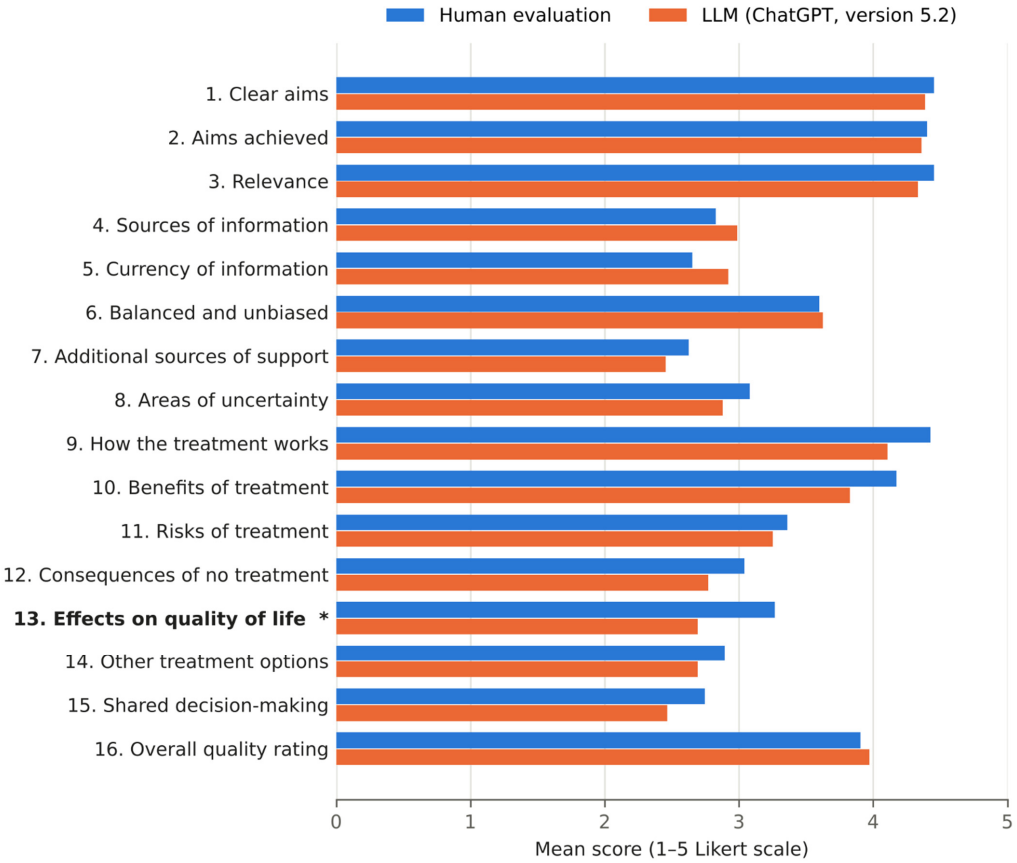


Figure 3. Comparison of DISCERN item scores. Comparison of DISCERN item scores between human evaluation and the LLM (ChatGPT, version 5.2) using a 5-point Likert scale. Items are labeled by content; asterisks indicate items that remained significant after Holm correction for multiple comparisons (Wilcoxon signed-rank test).

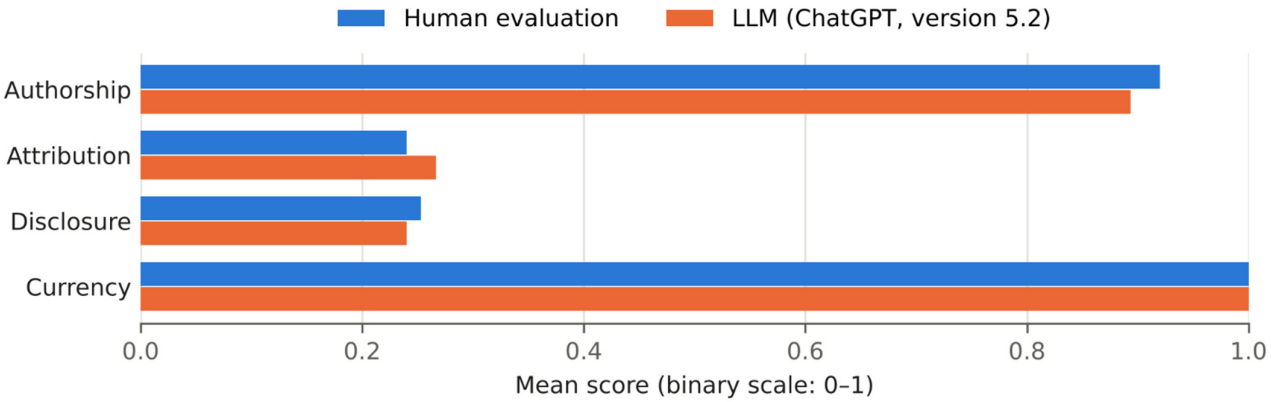


Figure 4. Comparison of JAMA benchmark criteria scores. Comparison of JAMA benchmark criteria scores between the human evaluation and the LLM (ChatGPT, version 5.2). Criteria include authorship, attribution, disclosure, and currency, each scored as present (1) or absent (0). No criterion differed significantly between the two methods (Wilcoxon signed-rank test).

4. Discussion

Total scores for the Educational Content Checklist, DISCERN, GQS, and JAMA criteria showed no statistically significant differences between human experts and the LLM. However, the absence of a statistically significant difference in total scores does not indicate that the two methods are equivalent; no equivalence analysis was performed. Moreover, the poor-to-moderate inter-method agreement shows that similar total scores do not imply high agreement in the evaluation of individual videos. These findings partly align with previous studies. ChatGPT-derived scores correlated with expert DISCERN and JAMA ratings for tonsillectomy videos [9], and no significant differences in DISCERN-5 or GQS scores were found for urological cancer videos, although repeatability was low [11]. In contrast, for fall-prevention exercise videos, GQS scores were similar, but ChatGPT assigned significantly higher DISCERN scores than the experts [10].

The inter-method ICC values, which ranged from poor (DISCERN and GQS) to moderate (ECC and JAMA), indicate that LLM ratings differed from expert ratings at the individual-video level, particularly for subjective and patient-centered criteria. Previous studies have likewise reported discrepancies in specific instruments despite similar overall scores [10] and low repeatability of ChatGPT evaluations [11]. Collectively, although LLMs might eventually assist with broad quality screening—a use that was not directly tested in this study—these findings do not support replacing human evaluation in critical educational and clinical safety assessments.

Importantly, aggregate score similarity does not imply flawless LLM performance. Goodman et al. reported that although median accuracy was high, only 39.4% of chatbot responses received the highest accuracy score, and accuracy correlated only modestly with completeness [18]. Broader systematic reviews of ChatGPT's reliability for health-related questions have similarly reported inconsistent accuracy across studies and underscored the need for cautious, supervised use in clinical settings [19]. Furthermore, a meta-analysis in radiology reported substantial variability in GPT-based diagnostic accuracy across models and studies [20]. Consequently, the absence of a significant difference in total scores should not be interpreted as equivalence to human expertise in complex or item-specific clinical contexts.

In the item-level analysis of the Educational Content Checklist, no statistically significant differences were detected between the LLM and human experts for contraindications, sterility and preparation, sonoanatomy, probe handling, local anesthetic spread, pharmacology and dosing, and complication management ($p > 0.05$). In contrast, significantly lower scores were assigned by the LLM for vascular safety (Doppler use) and needle technique.

For the appropriateness of indications, however, the difference was in the opposite direction, with the LLM assigning significantly higher scores than human experts ($p = 0.001$). The lower LLM scores for Doppler use and needle technique may be attributable to its reliance on static text and a limited number of screenshots rather than continuous video. Because the two methods did not receive identical inputs, these differences cannot be attributed solely to the capabilities of the model; these item-level findings should also be regarded as exploratory. A comparable study in radiology demonstrated that ChatGPT-4o's diagnostic accuracy was lowest when only static images were provided and improved substantially once clinical context was added, indicating that the model draws more diagnostic value from structured textual input than from static visual input alone [21]. This may parallel our findings: because dynamic, motion-dependent cues are difficult to convey through static screenshots, the LLM may have had less informative input in the domains requiring temporal interpretation; however, the independent effects of the different input types were not measured in this study, so this explanation remains a hypothesis. Thus, no statistically significant differences were detected for items related to static anatomical and theoretical content, whereas the LLM assigned lower scores for technical details that require interpretation of motion; because the accuracy of anatomical recognition was not tested directly, these findings do not demonstrate that the LLM correctly identified these structures.

In DISCERN-based evaluations, no statistically significant differences between the LLM and human experts were detected for most items, but the LLM assigned lower scores for patient-centered domains, including treatment mechanisms, benefits, consequences of non-treatment, effects on quality of life, and shared decision-making ($p < 0.05$). However, after Holm correction, only the difference for effects on quality of life (Item 13) remained significant, and these findings should be regarded as exploratory. This discrepancy may partly reflect the model's reliance on information explicitly available in the transcripts and selected screenshots; however, the contribution of these input characteristics could not be determined in the present study. As technical descriptions predominated in the videos, indirect or implicit patient-centered information may not have been adequately captured, resulting in lower scores. In addition, DISCERN was originally developed for written consumer information on treatment choices, and items addressing the consequences of no treatment, effects on quality of life, and shared decision-making are of limited relevance to professional procedural videos; this mismatch between instrument and content may also have contributed to the differences observed in these domains.

In contrast, human experts are able to integrate implicit contextual cues—such as publication date, channel credibility, target audience, and clinical context—into their assessments based on professional experience. Previous studies have reported that ChatGPT-generated medical information may vary in accuracy and completeness and may omit clinically relevant information [18,22]. Taken together, these findings suggest that in instruments centered on patient education and decision-making, such as DISCERN, LLM-based evaluations may exhibit limited contextual sensitivity. Enhancing the reliability of LLM-based assessments in these domains may require models to be trained not only on technical content but also on patient communication, clinical decision-making, and medical curriculum-based data.

In the present study, no statistically significant differences were observed between human experts and the LLM with respect to the JAMA benchmark criteria. This finding suggests that, for explicitly defined and objectively verifiable metadata, total LLM scores did not differ significantly from expert scores; however, agreement was only moderate (ICC = 0.529, 95% CI 0.343–0.674), and this does not establish that the LLM can substitute for expert assessment. Given that JAMA criteria rely largely on transparent information directly available within the content, this may explain why total scores did not differ significantly in this domain. In a related context, Walker et al. assessed ChatGPT-generated medical

information with the modified EQIP tool and found only 60% agreement with clinical guideline recommendations, despite fully consistent answers to repeated queries [22].

This study has several limitations. Restricting the analysis to five abdominal wall block techniques and to the 15 most-viewed English-language videos of 1–10 min with an available transcript for each technique limits the generalizability of the findings; therefore, the results should not be generalized to all educational videos on YouTube. The LLM-based evaluation relied on transcripts and limited static images rather than full-length videos, potentially constraining assessment of dynamic procedural elements. Because experts and the LLM did not receive identical inputs, the observed differences cannot be attributed solely to the limitations of the model; they may also reflect missing dynamic information in the inputs provided to the model. In addition, the independent contributions of transcripts, metadata, and procedural screenshots to the LLM evaluations were not investigated. Human–LLM comparisons using identical inputs and ablation analyses comparing transcript-only, screenshot-only, metadata-only, and combined-input conditions, as well as varying numbers of screenshots, should be addressed in future research. Transcripts were used without manual correction and may therefore contain errors in technical terminology. Although screenshots were selected according to predefined categories by an anesthesiologist who was not involved in the expert scoring, the selection process involved human judgment, and screenshot timestamps were not recorded. In addition, because each video was evaluated only once via the web interface, run-to-run variability could not be assessed. Additionally, the relatively small sample size may have reduced the ability to detect small differences, and the study was not designed to test equivalence; moreover, YouTube's dynamic search algorithms may yield different video results over time. Another limitation is that the Educational Content Checklist (ECC), although developed from published guidelines, the previous literature, and expert consensus to evaluate procedure-specific educational content, has not undergone formal psychometric validation. Although inter-rater reliability between the two human experts was high (ICC = 0.837, 95% CI 0.718–0.903), this reflects the consistency of its application rather than evidence of construct validity. Therefore, ECC-based findings should be interpreted with caution until the instrument is formally validated in independent datasets and different educational settings. In addition, the authors who developed the ECC also performed the expert scoring, which may have introduced familiarity bias in favor of the expert evaluations. DISCERN was designed for consumer information on treatment choices and may not fully suit professional procedural videos. Item-level analyses were exploratory and, despite Holm correction, should be interpreted with caution. Finally, the potential use of LLMs for pre-screening was not directly evaluated; no screening threshold was defined, and sensitivity, false-negative rates, processing time, and cost were not assessed. Future studies should evaluate these performance measures using predefined screening thresholds.

5. Conclusions

In this sample of the most-viewed English-language ultrasound-guided abdominal wall block training videos with available transcripts on YouTube, no significant differences were observed between experienced anesthesiologists and the LLM, which assessed transcripts and selected screenshots rather than the full videos, in overall quality and reliability scores. Although no statistically significant difference was found in total scores, this finding should not be interpreted as evidence of equivalence or interchangeability; agreement between the human experts and the LLM was poor or moderate depending on the instrument. In exploratory item-level analyses, score differences were observed in clinically important safety-related domains, particularly the Doppler-based vascular safety and needle technique, where the LLM assigned lower scores than experts; for the appropriateness of indications, the model diverged in the opposite direction by scoring more generously than experts.

Under the conditions of this study, the findings do not support substituting the LLM for expert human judgment in decisions related to procedural training and patient safety. LLM-assisted pre-screening combined with expert oversight is a potential application that should be tested prospectively, with predefined thresholds and assessment of sensitivity, false-negative rates, time, and cost.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162110650/s1, File S1: Standardized Evaluation Instructions for the Artificial Intelligence Model; File S2: Structured Educational Checklist (10 Items); File S3: DISCERN Instrument (16 Items); File S4: Global Quality Scale (GQS); File S5: JAMA Benchmark Criteria (4 Items).

Author Contributions: Conceptualization, N.A.E. and E.K.; methodology, N.A.E. and E.K.; formal analysis, E.K.; investigation, N.A.E. and E.K.; writing—original draft preparation, E.K.; writing—review and editing, N.A.E. and E.K. 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. This study exclusively analyzed publicly available YouTube videos and associated metadata and did not involve human participants, animals, clinical interventions, or identifiable personal data; therefore, institutional ethics committee approval was not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study, including the list of evaluated YouTube videos and the corresponding scoring data, are available from the corresponding author on reasonable request. The full prompt used for the LLM evaluation is provided in Supplementary File S1.

Acknowledgments: The authors thank Cansu Çiftçi, (Department of Anesthesiology and Reanimation, Konya City Hospital, Konya, Turkey), for selecting the screenshots and submitting the transcripts and screenshots to the large language model.

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

References

1. Rapp, A.K.; Healy, M.G.; Charlton, M.E.; Keith, J.N.; Rosenbaum, M.E.; Kapadia, M.R. YouTube is the most frequently used educational video source for surgical preparation. *J. Surg. Educ.* **2016**, *73*, 1072–1076. [CrossRef]
2. Curran, V.; Simmons, K.; Matthews, L.; Fleet, L.; Gustafson, D.L.; Fairbridge, N.A.; Xu, X. YouTube as an educational resource in medical education: A scoping review. *Med. Sci. Educ.* **2020**, *30*, 1775–1782. [CrossRef] [PubMed]
3. Alsharari, A.F.; Abuadas, F.H.; Alnassrallah, Y.S.; Salihu, D. Transversus abdominis plane block as a strategy for effective pain management in patients undergoing laparoscopic cholecystectomy: A systematic review. *J. Clin. Med.* **2022**, *11*, 6896. [CrossRef] [PubMed]
4. Sites, B.D.; Spence, B.C.; Gallagher, J.D.; Wiley, C.W.; Bertrand, M.L.; Blike, G.T. Characterizing novice behavior associated with learning ultrasound-guided peripheral regional anesthesia. *Reg. Anesth. Pain Med.* **2007**, *32*, 107–115. [CrossRef] [PubMed]
5. Tulgar, S.; Selvi, O.; Serifsoy, T.E.; Senturk, O.; Ozer, Z. YouTube as an information source of spinal anesthesia, epidural anesthesia and combined spinal-epidural anesthesia. *Rev. Bras. Anestesiol.* **2017**, *67*, 493–499. [CrossRef]
6. De Cassai, A.; Correale, C.; Sandei, L.; Ban, I.; Selvi, O.; Tulgar, S. Quality of erector spinae plane block educational videos on a popular video-sharing platform. *Cureus* **2019**, *11*, e4204. [CrossRef] [PubMed]
7. Tewfik, G.L.; Work, A.N.; Shulman, S.M.; Discepola, P. Objective validation of YouTube educational videos for the instruction of regional anesthesia nerve blocks: A novel approach. *BMC Anesthesiol.* **2020**, *20*, 168. [CrossRef] [PubMed]
8. Cerci Akcay, H.; Kargu, E.C.; Seker, N. Assessing the reliability and educational value of YouTube videos on computer-controlled local anesthesia in dentistry. *PLoS ONE* **2025**, *20*, e0329291. [CrossRef] [PubMed]
9. Serifler, S.; Gul, F. Evaluating tonsillectomy-related YouTube videos via a human expert review and the ChatGPT-4: A multi-method quality analysis. *BMC Med. Educ.* **2025**, *25*, 1157. [CrossRef] [PubMed]
10. Sozlu, U.; Kayiran, M.E.; Gunay, S.M.; Ozkazanc, G.; Basar, S. Evaluation of YouTube exercise videos for fall prevention in older adults: ChatGPT-4.5 versus human experts. *Geriatr. Bilim. Derg.* **2025**, *8*, 203–213. [CrossRef]

11. Demirci, A. A comparison of ChatGPT and human questionnaire evaluations of the urological cancer videos most watched on YouTube. *Clin. Genitourin. Cancer* **2024**, *22*, 102145. [CrossRef] [PubMed]
12. Taskin, K.; Yilmaz Ak, H. Artificial intelligence vs. human evaluation of anesthesia education videos: A comparative analysis using validated quality scales. *Front. Med.* **2026**, *13*, 1752664. [CrossRef] [PubMed]
13. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. Correction in *J. Chiropr. Med.* **2017**, *16*, 346.. [CrossRef] [PubMed]
14. Charnock, D.; Shepperd, S.; Needham, G.; Gann, R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health* **1999**, *53*, 105–111. [CrossRef] [PubMed]
15. Bernard, A.; Langille, M.; Hughes, S.; Rose, C.; Leddin, D.; Veldhuyzen van Zanten, S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **2007**, *102*, 2070–2077. [CrossRef]
16. Silberg, W.M.; Lundberg, G.D.; Musacchio, R.A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor—Let the reader and viewer beware. *JAMA* **1997**, *277*, 1244–1245. [CrossRef]
17. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
18. Goodman, R.S.; Patrinely, J.R.; Stone, C.A.; Zimmerman, E.; Donald, R.R.; Chang, S.S.; Berkowitz, S.T.; Finn, A.P.; Jahangir, E.; Scoville, E.A.; et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw. Open* **2023**, *6*, e2336483. [CrossRef] [PubMed]
19. Beheshti, M.; Toubal, I.E.; Alaboud, K.; Almalaysha, M.; Ogundele, O.B.; Turabieh, H.; Abdalnabi, N.; Boren, S.A.; Scott, G.J.; Dahu, B.M. Evaluating the reliability of ChatGPT for health-related questions: A systematic review. *Informatics* **2025**, *12*, 9. [CrossRef]
20. Nguyen, D.; Bronson, I.; Chen, R.; Kim, Y.H. A systematic review and meta-analysis of GPT-based differential diagnostic accuracy in radiological cases: 2023–2025. *Front. Radiol.* **2025**, *5*, 1670517. [CrossRef] [PubMed]
21. Atakir, K.; Isin, K.; Tas, A.; Onder, H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: Influence of image, clinical data, and answer options on performance. *Diagn. Interv. Radiol.* **2025**, *32*, 555. [CrossRef] [PubMed]
22. Walker, H.L.; Ghani, S.; Kuemmerli, C.; Nebiker, C.A.; Müller, B.P.; Raptis, D.A.; Staubli, S.M. Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. *J. Med. Internet Res.* **2023**, *25*, e47479. [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.