MEHES JOURNAL (MEDICAL AND HEALTH SCIENCES JOURNAL) | ISSN:3023-5804
Arrival Date: 07/02/2026 | Published Date: 27/03/2026 | 2026 Vol: 4 Issue: 1 pp: 1-14 |
DOI: https://doi.org/10.5281/zenodo.19238117

ARTICLE TYPE: RESEARCH ARTICLE

AI-Based Quality Assessment of Ultrasound-Guided Upper Extremity Nerve Block Videos on YouTube: A Comparison of Speech-Text-Only and Image-Enhanced Approaches

YouTube'daki Ultrason Rehberliğinde Üst Ekstremite Sinir Bloğu Videolarının Yapay Zekâ Tabanlı Kalite Değerlendirmesi: Yalnızca Konuşma-Metin ve Görüntü Destekli Yaklaşımların Karşılaştırılması

Esma Karaarslan[1]*, Cansu Çiftci[2]

*[1]Konya City Hospital, Department of Anesthesiology and Reanimation, Konya, Turkey
esmaayvaz@gmail.com, ORCID: 0000-0002-3459-0243
[2]Konya City Hospital, Department of Anesthesiology and Reanimation, Konya, Turkey
drdevecicansu@gmail.com, ORCID: 0009-0009-0215-4023

ÖZET

Amaç: Ultrasonografi eşliğinde yapılan üst ekstremite sinir bloğu uygulamaları ileri düzey görsel-motor koordinasyon gerektirmektedir. Bu alanda YouTube yaygın bir eğitim kaynağı olarak kullanılmakla birlikte, mevcut içeriklerin eğitsel kalitesi oldukça heterojendir. Bu çalışmada, ultrasonografi eşliğinde yapılan üst ekstremite sinir bloğu eğitim videolarında yapay zekâ (YZ) tabanlı kalite değerlendirmesinin yalnızca konuşma-metin temelli değerlendirme (Grup 1) ile görüntü destekli değerlendirme (Grup 2) yaklaşımları karşılaştırılarak etkinliğinin analiz edilmesi amaçlanmıştır.

Materyal ve Metot: Tanımlayıcı kesitsel tasarıma sahip bu çalışmada, 10 Aralık 2025 tarihinde YouTube'dan seçilen toplam 80 ultrasonografi eşliğinde yapılan üst ekstremite sinir bloğu eğitim videosu analiz edilmiştir. Her video, ChatGPT-5.2 (OpenAI) kullanılarak iki farklı değerlendirme koşulunda incelenmiştir. Eğitsel kalite; 10 maddelik Eğitim İçeriği Kontrol Listesi, DISCERN değerlendirme aracı, Global Kalite Ölçeği (GQS) ve Journal of the American Medical Association (JAMA) ölçütleri kullanılarak değerlendirilmiştir.

Bulgular: Grup 2, tüm kalite değerlendirme araçlarında Grup 1'e kıyasla istatistiksel olarak anlamlı derecede daha yüksek puanlar elde etmiştir: Eğitim İçeriği Skoru (33,90 ± 6,25'e karşı 29,11 ± 6,66; $p < 0,001$), DISCERN (45,01 ± 10,83'e karşı 37,90 ± 8,62; $p < 0,001$), GQS [medyan 4 (4–5)'e karşı 3 (3–4); $p < 0,001$] ve JAMA ölçütleri [medyan 2,5 (2–3)'e karşı 0 (0–0); $p < 0,001$].

Tartışma ve Sonuç: Görüntü destekli yapay zekâ yaklaşımı, özellikle görsel ipuçlarına dayalı işlemsel maddelerde daha yüksek kalite puanları sağlamıştır. Bu bulgular, yalnızca konuşma-metin temelli değerlendirmelerin ultrasonografi eşliğinde yapılan sinir bloğu eğitiminde kritik prosedürel unsurları sistematik olarak gözden kaçırabileceğini düşündürmektedir.

Anahtar Kelimeler: Yapay Zekâ, ChatGPT, Brakiyal Pleksus Bloğu, YouTube, Eğitsel Kalite

ABSTRACT

Objective: Ultrasound-guided upper extremity nerve block procedures require advanced visuomotor coordination. Although YouTube is widely used as an educational resource in this field, the educational quality of available content remains highly heterogeneous. This study aimed to compare the effectiveness of artificial intelligence (AI)–based quality assessment using a speech-text-only evaluation (Group 1) versus an image-enhanced evaluation (Group 2) for ultrasound-guided upper extremity nerve block educational videos.

Material and Methods: In this descriptive cross-sectional study, a total of 80 ultrasound-guided upper extremity nerve block educational videos selected from YouTube on December 10, 2025 were analysed. Each video was evaluated under two conditions using ChatGPT-5.2 (OpenAI). Educational quality was assessed using a 10-item Educational Content Checklist, the DISCERN instrument, the Global Quality Scale (GQS), and the Journal of the American Medical Association (JAMA) benchmark criteria.

Results: Group 2 demonstrated statistically significantly higher scores than Group 1 across all quality assessment tools: Educational Content Score (33.90 ± 6.25 vs. 29.11 ± 6.66; $p < 0.001$), DISCERN (45.01 ± 10.83 vs. 37.90 ± 8.62; $p < 0.001$), GQS [median 4 (4–5) vs. 3 (3–4); $p < 0.001$], and JAMA [median 2.5 (2–3) vs. 0 (0–0); $p < 0.001$].

Discussion and Conclusion: The image-enhanced AI approach yielded higher quality scores than the speech-text-only evaluation, particularly for procedurally dependent items reliant on visual cues. These findings suggest that speech-text-only evaluations may systematically fail to capture critical procedural elements in ultrasound-guided nerve block education.

Keywords: Artificial Intelligence, ChatGPT, Brachial Plexus Block, YouTube, Educational Quality

Sorumlu Yazar/Corresponding Author: Esma Karaarslan, Konya City Hospital, Department of Anesthesiology and Reanimation, Konya, Turkey, esmaayvaz@gmail.com, ORCID: 0000-0002-3459-0243

Atıf /Cite: Karaarslan E, Çiftci C. AI-Based Quality Assessment of Ultrasound-Guided Upper Extremity Nerve Block Videos on YouTube: A Comparison of Speech-Text-Only and Image-Enhanced Approaches. Mehes Journal. 27 Mart 2026;4(1):1-4.

INTRODUCTION

Ultrasound-guided peripheral nerve blocks are a cornerstone of contemporary regional anaesthesia practice and are widely used in upper extremity surgery through interscalene, supraclavicular, infraclavicular, and axillary approaches (1). These techniques provide effective perioperative analgesia but require advanced anatomical knowledge, refined hand–eye coordination, and a high level of visuomotor coordination for safe and effective performance (2). With the increasing integration of digital resources into medical education, YouTube has emerged as a commonly used platform for learning ultrasound-guided nerve block procedures among trainees and practitioners (3). Its open-access nature and visual format make it particularly attractive for procedural learning. However, as YouTube content is not subject to peer review, the accuracy, completeness, and educational quality of available videos may vary substantially (4–6). This variability raises concerns regarding the reliability of YouTube as an educational resource for technically demanding procedures.

To address these concerns, several structured evaluation instruments have been developed to assess the quality of online medical videos, including the DISCERN instrument, the Global Quality Scale (GQS), and the Journal of the American Medical Association (JAMA) benchmark criteria (4). While these tools provide standardized frameworks for assessment, evaluations conducted by human reviewers are inherently time-consuming and labour-intensive, limiting their feasibility for large-scale or continuous monitoring of educational content.

Recently, large language models (LLMs) have been explored as automated tools for evaluating health-related video content, offering the potential for scalable and efficient assessment (7). However, most previous studies have relied primarily on speech-text-only analyses, which may inadequately capture visually intensive elements of procedural education. For ultrasound-guided nerve blocks, critical aspects such as probe positioning, needle visualization, and real-time anatomical interpretation are inherently visual and dynamic, and the contribution of visual information to artificial intelligence (AI)–based quality assessment remains insufficiently explored.

Therefore, the present study aimed to compare AI–based evaluations of YouTube videos on ultrasound-guided upper extremity nerve blocks under two assessment conditions using the same AI model, ChatGPT (OpenAI, San Francisco, CA, USA; version GPT-5.2): a speech-text-only evaluation (Group 1) and an image-enhanced evaluation (Group 2).

MATERIALS AND METHODS

Study Design and Ethical Considerations

This study was conducted as a descriptive, cross-sectional content analysis comparing two AI–based evaluation approaches to assess the educational quality and information reliability of ultrasound-guided upper extremity nerve block videos on YouTube. All analyses were performed using the same AI model, ChatGPT-5.2, under two evaluation conditions: Group 1 and Group 2. The model was accessed via the official web-based interface, and default parameters were used throughout the study. All prompts were entered manually under controlled conditions to ensure consistency across evaluations.

Only publicly accessible video content and associated metadata were analysed. The study did not involve human participants, patient data, or clinical interventions; therefore, institutional review board approval and informed consent were not required. The study was conducted in accordance with the Declaration of Helsinki.

Video Search Strategy and Selection Process

A systematic search of the YouTube platform was performed on December 10, 2025. To minimise the influence of personalised search algorithms, all searches were conducted in incognito browser mode with cookies and browsing history cleared. The following search terms were used sequentially: “interscalene block ultrasonography,” “supraclavicular block ultrasonography,” “infraclavicular block ultrasonography,” and “axillary block ultrasonography.”

For each search term, videos were sorted by view count in descending order, and the first 20 videos meeting predefined inclusion criteria were selected. Accordingly, a total of 80 videos (20 per block technique) were included in the analysis.

Inclusion and Exclusion Criteria

Videos were included if they provided educational or procedural content specifically focusing on ultrasound-guided interscalene, supraclavicular, infraclavicular, or axillary nerve blocks; had a duration between one and ten minutes; and featured English-language narration or a verifiable English transcript. Only original, full-length YouTube videos were considered for analysis. Conversely, videos were excluded if they were shorter than one minute or longer than ten minutes; lacked audio or transcriptions; were presented in a language other than English; or demonstrated a mismatch between the title and clinical content. Additionally, duplicate uploads and videos irrelevant to the specific procedural techniques were excluded from the final dataset.

Validation and Interobserver Agreement

To assess the accuracy and clinical relevance of the AI–based evaluations, a random subset of 20 videos was independently scored by two anaesthesiology specialists using the same

assessment tools. Interobserver agreement was analysed using the intraclass correlation coefficient (ICC), which demonstrated a high level of consistency between the evaluators.

Videos with discrepant scores between the two experts were subsequently reviewed by a third senior anaesthesiology specialist, and a final consensus score was established through joint evaluation.

AI–Based Evaluation

Each video was evaluated under two conditions using the same AI model to ensure methodological consistency. In Group 1, the complete video speech text was analysed without visual input.

In Group 2, the same speech text was evaluated together with visual inputs. The first image consisted, when available, of a screenshot displaying the channel and presenter information. This was followed by screenshots representing key procedural stages, including patient positioning, ultrasound probe placement, identification of anatomical structures, needle visualization, and local anaesthetic spread.

A total of at least five and at most ten images were used for each video. The number of images was determined according to the video duration and scene variation. Due to system limitations, a maximum of ten images could be uploaded per evaluation session.

All evaluations were performed using the same standardized prompt (Supplementary File 1), with input modality as the only variable.

Assessment Framework

Educational content quality and information reliability were assessed using a Structured Educational Content Checklist (Supplementary File 2), the DISCERN instrument (Supplementary File 3), GQS (Supplementary File 4), and JAMA benchmark criteria (Supplementary File 5). Item-level and total scores generated by the AI model were recorded separately for each group.

Video Characteristics and Quantitative Variables

For each video, duration, channel subscriber count, time since upload, number of views, likes, and comments were recorded. Engagement rate was calculated as [(likes + comments) / views] × 100, and view rate was calculated as total views divided by the number of days since upload.

Statistical Analysis

All analyses were performed using IBM SPSS Statistics version 20.0 (IBM Corp., Chicago, IL, USA). Continuous variables are presented as mean ± standard deviation or median (25th–

75th percentiles), as appropriate, and categorical variables as frequencies and percentages. Data normality was assessed using the Kolmogorov–Smirnov test.

Paired comparisons between the two evaluation conditions were conducted using the paired-samples t-test for normally distributed data and the Wilcoxon signed-rank test for non-normally distributed data. Associations between video characteristics and quality scores were examined using Pearson or Spearman correlation coefficients, as appropriate.

All tests were two-tailed, and $p < 0.05$ was considered statistically significant.

RESULTS

The general characteristics of the included videos are summarised in Table 1. While video durations were relatively comparable due to the predefined inclusion criteria, channel popularity as well as viewing and engagement metrics showed wide variability.

Table 1. General characteristics of the videos.

Variable	Median (25th–75th Percentiles)	Minimum–Maximum Value
Subscriber count	11,350 (3,005–72,450)	106–2,550,000
Video duration (minutes)	4.78 (2.77–6.20)	1.23–9.52
Upload time (months)	70.50 (47–153)	6–204
Like count	217.50 (24.50–517)	0–22,000
View count	38,260.00 (5,545.50–99,804.00)	262–414,188
Comment count	3 (1–13)	0–143
Engagement rate (%)	0.69 (0.28–1.10)	0–11.28
View rate (views/day)	15.22 (2.79–38.11)	0.09–242.22

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

Interobserver reliability of the assessment tools used in the study was high across all scales. The ICC was 0.957 for the Educational Content Score, 0.950 for DISCERN, 0.927 for GQS, and 0.816 for JAMA ($p < 0.001$ for all scales).

In analyses comparing the performance of the AI model with expert ratings, statistically significant agreement with expert scores was observed across all assessment tools in Group 2, which used the image-enhanced evaluation approach. For Group 2, ICC values were 0.956 for the Educational Content Score and 0.968 for DISCERN ($p < 0.001$ for both). In contrast, Group 1, which relied on the speech-text-only evaluation, demonstrated lower ICC values for the same scales.

For the JAMA benchmark criteria, Group 1 did not show statistically significant agreement with expert ratings (ICC = 0.001; $p = 0.495$). Conversely, when visual metadata were incorporated in Group 2, JAMA scores demonstrated statistically significant agreement with expert ratings (ICC = 0.876; $p < 0.001$).

In paired comparisons of the same videos, Group 2 yielded significantly higher scores than Group 1 across all quality metrics (Table 2; all $p < 0.001$).

Table 2. Quality evaluation scores under two evaluation conditions for the same videos (Group 1 vs. Group 2).

	Group 1	**Group 2**		
Video Evaluation Parameters	**Value**	**Value**	**Paired difference (95% CI)**	***P* Value**
Educational Content Score	29.11±6.66	33.90±6.25	−4.79 (−6.80 to −2.77)	<0.001
DISCERN Score	37.90±8.62	45.01±10.83	−7.11 (−10.17 to −4.06)	<0.001
Global Quality Score (GQS)	3 (3–4)	4 (4–5)	-1.0 (-1.0 to 0.0)	<0.001
JAMA Score	0 (0–0)	2.5 (2–3)	-2.0 (-3.0 to -2.0)	<0.001

Data are presented as mean ± standard deviation for normally distributed variables and as median (25th–75th percentiles) for non-normally distributed variables. Paired differences are presented as mean paired differences with 95% confidence intervals (CI) for normally distributed variables and as Hodges–Lehmann median paired differences with 95% CI for non-normally distributed variables. Comparisons between Group 1 and Group 2 were performed using the paired-samples t-test or the Wilcoxon signed-rank test, as appropriate. Statistical significance was set at $p < 0.05$.

As shown in Table 3, Group 2 achieved significantly higher scores than Group 1 in visually dependent domains, including Indications and Positioning, Sterility and Preparation, Sonoanatomy, Vascular Safety/Doppler, Probe Handling and Image Optimisation, Needle Technique, and Local Anaesthetic Spread/Hydrodissection (all $p < 0.05$).

In contrast, no statistically significant differences were observed between the groups for the items pharmacology and dosage ($p = 0.202$) and complication management ($p = 0.465$).

Table 3. Distribution of Responses to the Educational Content Checklist Items.

			Group 1	**Group 2**	
Item	Score	Category	n (%)	n (%)	*p* value
Indications & Positioning	1	No information	7 (8.8%)	2 (2.5%)	
	2	Very poor	14 (17.5%)	8 (10.0%)	<0.001
	3	Poor	12 (15.0%)	11 (13.8%)	
	4	Good	32 (40.0%)	25 (31.3%)	
	5	Very good	15 (18.8%)	34 (42.5%)	
Contraindications	1	No information	69 (86.3%)	62 (77.5%)	
	2	Very poor	9 (11.3%)	12 (15.0%)	
	3	Poor	1 (1.3%)	3 (3.8%)	0.014
	4	Good	0 (0.0%)	2 (2.5%)	
	5	Very good	1 (1.3%)	1 (1.3%)	
Sterility & Preparation	1	No information	59 (73.8%)	15 (18.8%)	
	2	Very poor	12 (15.0%)	22 (27.5%)	
	3	Poor	4 (5.0%)	18 (22.5%)	<0.001
	4	Good	3 (3.8%)	21 (26.3%)	
	5	Very good	2 (2.5%)	4 (5.0%)	
Sonoanatomy	1	No information	2 (2.5%)	0 (0.0%)	
	2	Very poor	0 (0.0%)	1 (1.3%)	
	3	Poor	9 (11.3%)	2 (2.5%)	<0.001
	4	Good	32 (40.0%)	8 (10.0%)	
	5	Very good	37 (46.3%)	69 (86.3%)	
Vascular Safety (Doppler)	1	No information	26 (32.5%)	18 (22.5%)	
	2	Very poor	12 (15.0%)	14 (17.5%)	0.003
	3	Poor	16 (20.0%)	14 (17.5%)	
	4	Good	13 (16.3%)	12 (15.0%)	
	5	Very good	13 (16.3%)	22 (27.5%)	
Probe & Image Optimisation	1	No information	7 (8.8%)	0 (0.0%)	
	2	Very poor	8 (10.0%)	1 (1.3%)	<0.001
	3	Poor	19 (23.8%)	6 (7.5%)	
	4	Good	23 (28.7%)	32 (40.0%)	
	5	Very good	23 (28.7%)	41 (51.2%)	
Needle Technique	1	No information	3 (3.8%)	5 (6.3%)	
	2	Very poor	7 (8.8%)	1 (1.3%)	
	3	Poor	12 (15.0%)	8 (10.0%)	<0.001
	4	Good	32 (40.0%)	20 (25.0%)	
	5	Very good	26 (32.5%)	46 (57.5%)	
LA Spread (Hydrodissection)	1	No information	10 (12.5%)	8 (10.0%)	
	2	Very poor	7 (8.8%)	3 (3.8%)	<0.001
	3	Poor	11 (13.8%)	8 (10.0%)	
	4	Good	28 (35.0%)	18 (22.5%)	
	5	Very good	24 (30.0%)	43 (53.8%)	
Pharmacology & Dose	1	No information	30 (37.5%)	25 (31.3%)	
	2	Very poor	12 (15.0%)	12 (15.0%)	
	3	Poor	12 (15.0%)	13 (16.3%)	0.202
	4	Good	18 (22.5%)	18 (22.5%)	
	5	Very good	8 (10.0%)	12 (15.0%)	
Complication Management	1	No information	27 (33.8%)	25 (31.3%)	
	2	Very poor	18 (22.5%)	17 (21.3%)	0.465
	3	Poor	19 (23.8%)	20 (25.0%)	
	4	Good	12 (15.0%)	13 (16.3%)	
	5	Very good	4 (5.0%)	5 (6.3%)	

Data are presented as n (%). Group comparisons were performed using the Wilcoxon signed-rank test, as the same videos were evaluated under two conditions. Statistical significance was set at $p < 0.05$.

In Group 1, video duration showed moderate positive correlations with the Educational Content Score, DISCERN, and GQS (all $p < 0.01$) (Table 4). View rate demonstrated weak

but significant correlations with these scores (all $p < 0.05$), whereas engagement rate showed a weak association only with GQS ($p < 0.05$). Subscriber count was not significantly correlated with any quality metric ($p > 0.05$).

Strong correlations were observed among the primary quality scores (all $p < 0.001$). In contrast, JAMA scores were not significantly associated with the other quality measures in Group 1 ($p > 0.05$).

Table 4. Correlations between video characteristics and quality scores in Group 1.

Values are presented as correlation coefficients (r). Pearson or Spearman correlation tests were applied as appropriate. Statistical significance was set at $p < 0.05$.

	Educational Content Score (*r*)	***p***	**DISCERN Score (*r*)**	***p***	**GQS (*r*)**	***p***	**JAMA Score (*r*)**	***p***
Subscriber Count	0.169	0.135	0.160	0.157	0.192	0.089	-0.068	0.551
Video duration (minutes)	0.392	<0.001	0.389	<0.001	0.335	0.002	0.136	0.231
Engagement Rate	0.025	0.825	0.161	0.154	0.247	0.028	-0.178	0.114
View rate	0.277	0.013	0.240	0.032	0.290	0.009	-0.059	0.603
DISCERN score	0.635	<0.001	—	—	0.750	<0.001	-0.027	0.814
GQS score	0.722	<0.001	0.750	<0.001	—	—	0.036	0.751
JAMA score	0.108	0.340	-0.027	0.814	0.036	0.751	—	—

In Group 2, broader associations were observed between video characteristics and quality scores (Table 5). Subscriber count showed weak positive correlations with the Educational Content Score, DISCERN, GQS, and JAMA (all $p < 0.05$). Video duration and view rate also demonstrated weak but significant correlations with the Educational Content Score, DISCERN, and GQS (all $p < 0.05$).

Inter-metric analysis revealed strong correlations between the Educational Content Score and DISCERN, GQS, and JAMA, as well as between DISCERN and GQS and JAMA (all $p < 0.05$). JAMA scores were also significantly correlated with GQS ($p < 0.01$).

Table 5. Correlations between video characteristics and quality scores in Group 2.

	Educational Content Score (*r*)	***p***	**DISCERN Score (*r*)**	***p***	**GQS (*r*)**	***p***	**JAMA Score (*r*)**	***p***
Subscriber Count	0.233	0.037	0.251	0.025	0.255	0.022	0.384	<0.001
Video duration (minutes)	0.304	0.006	0.239	0.033	0.229	0.041	-0.062	0.588
Engagement Rate	0.110	0.330	0.310	0.005	0.278	0.012	0.076	0.503
View rate	0.289	0.009	0.228	0.042	0.223	0.047	0.340	0.002
DISCERN score	0.739	<0.001	—	—	0.781	<0.001	0.414	<0.001
GQS score	0.690	<0.001	0.781	<0.001	—	—	0.348	0.002
JAMA score	0.259	0.021	0.414	<0.001	0.348	0.002	—	—

Values are presented as correlation coefficients (r). Pearson or Spearman correlation tests were applied as appropriate. Statistical significance was set at $p < 0.05$.

In the 16-item DISCERN analysis (Figure 1), Group 2 demonstrated significantly higher scores across multiple domains. Within the publication reliability domain, image-enhanced evaluation yielded higher scores for clarity and achievement of aims, source reporting, currency, balanced presentation, and provision of additional information (all $p < 0.05$). Similarly, in the treatment information domain, Group 2 scored significantly higher for explanation of treatment mechanisms and benefits, consequences of no treatment, impact on quality of life, presentation of alternative options, and support for shared decision-making (all $p < 0.05$). No significant differences were observed for relevance to the topic, discussion of uncertainty, or reporting of treatment risks ($p > 0.05$). The overall DISCERN assessment score was also significantly higher in Group 2 ($p < 0.001$).

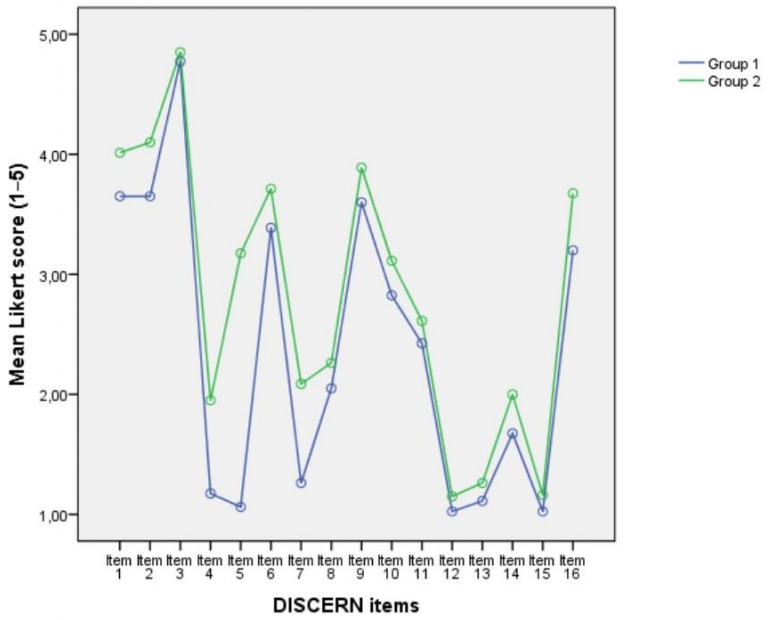

Figure 1. Mean DISCERN item scores by group (Group 1: Speech-Text-Only; Group 2: Image-Enhanced).

The group-based comparison of mean scores for the Journal of the American Medical Association (JAMA) benchmark criteria—authorship, attribution, transparency, and currency—is presented in Figure 2. In paired comparisons of scores obtained from the same videos, Group 2 demonstrated statistically significantly higher scores across all JAMA components compared with Group 1 ($p < 0.001$ for all parameters).

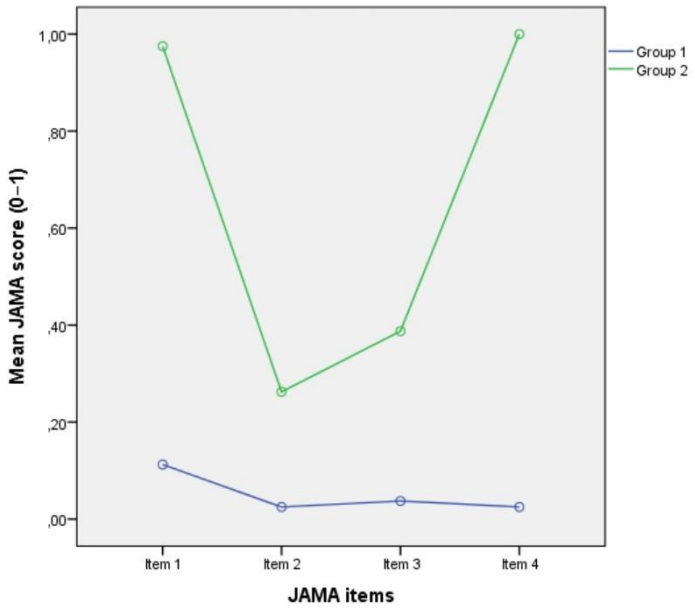

Figure 2. Mean JAMA item scores by group. (Group 1: Speech-Text-Only; Group 2: Image-Enhanced).

DISCUSSION

This study demonstrates that incorporating visual inputs into a speech-text-only evaluation framework significantly alters AI–generated quality scores for YouTube videos on ultrasound-guided upper extremity nerve blocks. The largest differences were observed in domains inherently dependent on visual interpretation, including indications and positioning, sterility and preparation, sonoanatomy, probe placement and image optimisation, needle technique, and local anaesthetic spread. These findings indicate that the core educational components of ultrasound-guided regional anaesthesia are primarily conveyed through visual cues and that evaluations based solely on verbal narration may incompletely reflect procedural competence.

Our results are consistent with previous studies showing that visually rich procedural videos are interpreted more accurately by large language models than text-dominant content (8,9). Visual context appears to enhance LLM-based interpretation, underscoring the importance of image-enhanced input when evaluating visually intensive procedural education materials.

The image-enhanced evaluation approach also resulted in higher scores across DISCERN subdomains related to source identification, information currency, and presentation of alternative options. This likely reflects the fact that references, institutional affiliations, and visual indicators of alternative approaches are often presented on-screen but are not consistently captured in speech text. Similarly, JAMA benchmark scores for authorship, attribution, transparency, and currency improved when visual elements were included. Opening sequences, subtitles, on-screen text, and institutional logos provide key cues regarding content provenance and credibility, which speech-text-only evaluation is inherently limited in detecting. These findings align with recent work demonstrating more accurate identification of transparency and reliability metrics when all video components are analysed (7).

The performance gains observed with image-enhanced evaluation parallel expert-based frameworks for assessing procedural competence. In the Delphi-based tool developed by Cheung et al., visually anchored skills such as sonoanatomy recognition, needle visualisation, and procedural execution were identified as core competencies (10). The selective improvement of these domains in Group 2 suggests that, when provided with visual input, AI may approximate the visually oriented evaluative strategies used by human experts. Concurrent improvements in DISCERN and JAMA scores further support the role of visual credibility cues, including institutional branding and professional graphical elements, which humans commonly use to infer reliability in online environments (11).

Only weak associations were identified between video popularity metrics and educational quality. Under speech-text-only evaluation, view rate showed modest correlations with educational content, DISCERN, and GQS scores, while engagement rate was weakly associated only with GQS. With image-enhanced evaluation, broader—but still weak—relationships emerged between reach-related metrics and quality scores, particularly for JAMA. These findings suggest that popularity indicators do not reliably reflect educational quality, consistent with prior evidence that viewer preferences are often influenced by visual appeal or platform algorithms rather than informational depth (4,12).

The strong agreement between expert ratings and image-enhanced AI evaluations can be interpreted through cognitive load theory and dual-coding principles. Procedural skills such as needle tracking and anatomical orientation are spatially complex and cannot be fully represented through text alone. Coordinated visual and verbal information has been shown to enhance both learning and assessment effectiveness (13), which may explain the closer alignment between expert judgment and image-enhanced AI outputs observed in this study.

Several limitations should be acknowledged. Visual inputs were limited to selected screenshots rather than full video streams, potentially overlooking dynamic procedural elements. Analyses were performed using a single LLM, and expert validation was conducted on a subset of videos, which may limit generalisability. In addition, only English-language videos were included.

Future studies incorporating continuous video-stream analysis, multiple AI models capable of processing both text and visual inputs, and broader expert validation datasets are warranted to further clarify the role of image-enhanced AI in evaluating procedural medical education content.

CONCLUSION

This study shows that incorporating visual inputs into AI–based evaluation frameworks substantially influences the assessment of educational quality in YouTube videos on ultrasound-guided upper extremity nerve blocks. Image-enhanced evaluation demonstrated closer agreement with expert assessments, particularly for visually intensive procedural components that are central to ultrasound-guided regional anaesthesia education.

These findings indicate that speech-text-only evaluation approaches may be insufficient for assessing procedural medical education content and underscore the importance of AI systems that integrate visual information alongside speech text. While AI should not replace expert judgment, it may serve as a scalable and efficient adjunct for the preliminary evaluation, systematic screening, and quality assurance of online educational resources.

Future studies incorporating continuous video analysis, multiple AI models capable of processing both speech text and visual inputs, and broader expert validation are warranted to further refine the role of AI in evaluating procedural medical education materials.

Scientific Responsibility Statement

The authors declare that they are responsible for the scientific content of the article, including the study design, data collection, analysis and interpretation of data, drafting of the manuscript, scientific review of the content, and approval of the final version of the article.

Ethics Approval and Consent

This study analyzed publicly available online video content and associated metadata only and did not involve human participants, patient data, or clinical interventions. Therefore, institutional review board approval and informed consent were not required. In accordance with institutional policies, a formal ethics approval waiver was not applicable for this study.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

E.K. conceived and designed the study, performed the formal analysis, and drafted the manuscript. C.Ç. contributed to data curation, methodology, and critical revision of the manuscript. All authors approved the final version for submission.

Declaration on the Use of Artificial Intelligence in the Study and Writing Process

In this study, AI was used as the primary subject of investigation to systematically evaluate the educational content quality of YouTube videos using speech-text-only and image-enhanced evaluation approaches.

In addition, Grammarly and ChatGPT were used in a limited manner during the writing process solely to improve language clarity and readability. All scientific content, methodology, data analysis, results, and interpretations were generated and critically reviewed by the authors. No data were generated, analyzed, or modified using AI tools. The authors take full responsibility for the accuracy and integrity of the work.

Financial Support/Funding

There is no funding support for this study.

Acknowledgements

The authors have no acknowledgements to declare.

REFERENCES

1. Marhofer P, Chan VWS. Ultrasound-guided regional anesthesia: current concepts and future trends. Anesth Analg. 2007;104(5):1265-9. https://doi.org/10.1213/01.ane.0000260614.32794.7b
2. Sites BD, Brull R, Chan VW, Spence BC, Gallagher J, Beach ML, et al. Artifacts and pitfall errors associated with ultrasound-guided regional anesthesia: part II: a pictorial approach to understanding and avoidance. Reg Anesth Pain Med. 2007;32(5):419-33. https://doi.org/10.1016/j.rapm.2007.08.001
3. Rapp AK, Healy MG, Charlton ME, Keith JN, Rosenbaum ME, Kapadia MR. YouTube is the most frequently used educational video source for surgical preparation. J Surg Educ. 2016;73(6):1072-6. https://doi.org/10.1016/j.jsurg.2016.04.024
4. Drozd B, Couvillon E, Suarez A. Medical YouTube videos and methods of evaluation: literature review. JMIR Med Educ. 2018;4(1):e3. https://doi.org/10.2196/mededu.8527
5. Gupta T, Haidery TH, Sharma R, Sharma S, Kumar A. How reliable are YouTube videos for general surgery residents learning? Cureus. 2023;15(2):e34718. https://doi.org/10.7759/cureus.34718
6. Cho NR, Park S, Choi JB, Park J, Kim D, Park K. Reliability and quality of YouTube videos on ultrasound-guided brachial plexus block: a programmatical review. Healthcare (Basel). 2021;9(8):1083. https://doi.org/10.3390/healthcare9081083
7. Khalil M, Mohamed F, Shoufan A. Evaluating the quality of medical content on YouTube using large language models. Sci Rep. 2025;15:9906. https://doi.org/10.1038/s41598-025-94208-6
8. Serifler S, Gul F. Evaluating tonsillectomy-related YouTube videos via a human expert review and ChatGPT-4: a multi-method quality analysis. BMC Med Educ. 2025;25:1157. https://doi.org/10.1186/s12909-025-07739-x
9. Thakur N, Han CY, Huang Y, Singh AD. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26:e59505. https://doi.org/10.2196/59505
10. Cheung JJ, Chen EW, Darani R, McCartney CJ, Dubrowski A, Awad IT. The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. Reg Anesth Pain Med. 2012;37(3):329-33. https://doi.org/10.1097/AAP.0b013e318246f63c
11. Metzger MJ, Flanagin AJ. Credibility and trust of information in online environments: the use of cognitive heuristics. J Pragmat. 2013;59(Pt B):210-20. https://doi.org/10.1016/j.pragma.2013.07.012
12. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: a systematic review. Health Informatics J. 2015;21(3):173-94. https://doi.org/10.1177/1460458213512220
13. Mayer RE, Fiorella L, Stull A. Five ways to increase the effectiveness of instructional video. Educ Technol Res Dev. 2020;68(3):837-52. https://doi.org/10.1007/s11423-020-09749-6