Turk Neurosurg 36(2):200-210, 2026
DOI: 10.5137/1019-5149.JTN.49058-25.2

Original Investigation
Education in Neurosurgery

Received: 30.04.2025
Accepted: 05.08.2025
Published Online: 04.03.2026

Beyond Human 'Eyes' in Neurosurgical Exams: Success of Artificial Intelligence (ChatGPT-4o, Grok, and Gemini) in the Image-Based Questions of Turkish Neurosurgical Society Proficiency Board Exams

Alperen SOZER[1], Gokberk EROL[2], Ozan Yavuz TUFEK[3], Batuhan SOZER[4], Merve BUKE SAHIN[5], Mustafa Caglar SAHIN[6]

[1]Sincan Training and Research Hospital, Department of Neurosurgery, Ankara, Türkiye
[2]Adiyaman Training and Research Hospital, Department of Neurosurgery, Adiyaman, Türkiye
[3]Gazi University Faculty of Medicine, Department of Neurosurgery, Ankara, Türkiye
[4]Ankara Medipol University, Faculty of Medicine, Ankara, Türkiye
[5]Kulu District Health Directorate, Department of Public Health, Konya, Türkiye
[6]Kulu State Hospital, Department of Neurosurgery, Konya, Türkiye

Corresponding author: Mustafa Caglar SAHIN ✉ dr.mcaglarsahin@gmail.com

ABSTRACT

AIM: To evaluate the impact of generative artificial intelligence and large language models (LLMs) on medical training and neurosurgical education, specifically focusing on their emerging capabilities in image interpretation.

MATERIAL and METHODS: This study evaluated the performance of three major LLMs (ChatGPT-4o, Grok, and Gemini) on image-based neurosurgical proficiency board questions and compared their latest versions.

RESULTS: Real-life candidates answered correctly 70.75% of the time. LLMs answered correctly 47.38% of the time and were significantly outperformed by the candidates. Prompt selection was found to significantly influence the performance of GPT and Grok, but not Gemini. Matching and significantly outperforming the candidates was only possible by combining the best answers from all three LLMs across four runs.

CONCLUSION: Although previous research has demonstrated strong capabilities of LLMs in text-only questions, this the results of the present study revealed that image analysis abilities of these models need further improvement when compared to actual candidates. Furthermore, the impact of prompt selection and repeated questioning should be emphasized, particularly when seeking correlation with the real-life exam results.

KEYWORDS: Artificial Intelligence, ChatGPT-4o, Education, Gemini, Grok

Alperen SOZER : 0000-0001-6475-7094
Gokberk EROL : 0000-0001-6651-5486
Ozan Yavuz TUFEK : 0000-0002-8157-8829
Batuhan SOZER : 0000-0002-3312-3475
Merve BUKE SAHIN : 0000-0002-5132-8220
Mustafa Caglar SAHIN : 0000-0002-5141-8154

INTRODUCTION

Driven by technological advancements that revolutionize how medicine is taught and proficiency is assessed, medical education is continuously evolving. With the rapid advancement of artificial intelligence (AI), large language models (LLMs) have emerged as pivotal tools in various domains, including healthcare. These models not only assist in decision making, but also can transform traditional educational and assessment methods. Since medicine prioritizes diagnostic accuracy and problem-solving skills, the integration in this field of LLMs such as ChatGPT, Grok, and Gemini represents a transformative step. In this context, the present study seeks to explore this paradigm shift within the context of neurosurgical board examinations, with a particular emphasis on the importance of image-based evaluation in clinical scenarios.

ChatGPT-4o is an advanced language model developed to interact and assist users in a conversational manner (13). Building upon its predecessors, ChatGPT-4o combines enhanced reasoning, problem-solving capabilities, and contextual understanding. Its multimodal abilities, including text and image input analysis—referred to by the recently added letter 'o', which stands for 'omni'—have positioned ChatGPT-4o as a versatile tool not only in daily queries, but also in specialized domains such as medicine, engineering, and education.

Furthermore, on November 4, 2023, xAI, based in San Francisco, introduced Grok, an innovative artificial intelligence model. Unlike existing large language models, Grok integrates "real-time knowledge" from the X platform (formerly Twitter), enabling it to provide users with the most current information (20).

In its turn, Google Gemini, developed by Google DeepMind and introduced in December 2023, represents a significant advancement in artificial intelligence as a multimodal large language model. Capable of processing diverse data types such as text, images, audio, video, and code, Gemini expands the boundaries of generative AI. To date, it has been seamlessly integrated into products like Bard, thus effectively supporting users in tasks such as writing, planning, and learning. Its advanced language comprehension and generation capabilities position it as a strong competitor to models like ChatGPT, highlighting its potential in transforming AI applications across various scientific and practical domains (6).

While previous research demonstrated ChatGPT's notable success in answering textual questions in the Turkish Neurosurgical Society Proficiency Board Exams (TNSPBE), with an accuracy significantly exceeding that of human candidates (16), an important limitation that remains is the exclusion of questions containing visual elements, which constitute a substantial and clinically relevant portion of neurosurgical evaluations. This omission is due to the earlier incapability of ChatGPT to process and interpret images.

Since visual interpretation is now a key feature of major LLMs, investigating this aspect becomes essential. To fill this gap in the literature, the present study was conducted to evaluate this new feature and to compare the three major LLMs used nowadays.

MATERIAL and METHODS

This study evaluated questions containing visual elements from the last eight written TNSPBE exams. The exam consisted of multiple-choice questions, each with 5 answer options. The questions and answer keys are publicly available on the Turkish Neurosurgical Society website. Candidate responses for all questions were obtained in an anonymized format, upon obtaining the permission of the Turkish Neurosurgery Association Board of Directors. No additional approval was required for this investigation by local and regional regulations. Analyses were conducted on a total of 108 questions containing visual elements. The distribution of image types and content areas of the investigated questions is summarized in Table I.

The questions were submitted to three major LLMs (namely, ChatGPT-4o, Grok-vision-beta, Gemini-1.5-pro) using their respective API services between November 25, 2024, and November 30, 2024. To the best of our knowledge, none of these services received major updates during this period. The builds used were as follows: GPT-4o (gpt-4o-2024-08-06, released August 6, 2024), Gemini (gemini-1.5-pro-002, released September 24, 2024), and Grok (grok-vision-beta-0.1.0, released November 23, 2024).

Visual components were manually extracted from the exam questions in their original layout and were saved, as close as possible to their original printed resolution and color settings, as .jpg files. The text parts of the questions were extracted and stored in text format. Python was used for all remaining applications. Next, the extracted images were converted to base64 format using the Python standard library (2). OpenAPI SDK (14) was used to send prompts to all LLMs.

More specifically, two different prompts—referred to as the NS prompt and the HQ prompt—were used, and each question was sent to each LLM twice for each of these prompts (i.e., a total of four submissions per question). The NS-prompt instructed the LLM to assume the role of a neurosurgeon, while the HQ-prompt framed the question as part of a standard quiz. The prompts were displayed in Code 1. All responses were then recorded and inspected.

Statistical Analysis

For the statistical analysis, Cochran's Q test was used to compare the correct answer rates of all three LLMs simultaneously, while McNemar's test with Bonferroni correction was applied for pairwise comparisons between LLMs. Asymptotic significances were evaluated using χ^2 statistics when the sample size adequacy assumption was met; otherwise, exact *p*-values were reported. Two proportion z-tests were used to compare the correct answer rates of LLMs with those of human candidates. The associations between LLMs' correct answer statuses and candidate performance were evaluated using Spearman correlation. Correlation coefficients interpreted according to the previously established cut-off values (4). All results were evaluated at a 95% confidence interval with a

Table I: Distribution of Image Modalities Across Different Neurosurgical Content Areas. The Table Presents the Number of Questions Categorized by Anatomical and Clinical Areas, Along with Their Corresponding Imaging Types

	Anatomy	Paediatric	Cranial	Spinal	Trauma	Vascular	Total
MR	1	13	16	6	0	1	37
Cadaver	19	0	0	0	0	0	19
Multimodal	1	2	1	5	4	2	15
CT	0	4	1	1	4	1	11
Illustration	7	0	0	2	0	1	10
Real Picture	1	6	0	0	0	0	7
X-Ray	1	1	0	3	0	0	5
DSA	0	0	0	0	0	3	3
Micrograph	0	0	0	1	0	0	1
Total	30	26	18	18	8	8	108

MR: *Magnetic resonance.* ***CT:*** *Computerized tomography.* ***DSA:*** *Digital subtraction angiography.*

significance level of $p<0.05$. All statistical analyses were conducted using SPSS® Statistics (version 27, IBM® Corp., USA).

RESULTS

Of the 4,988 candidate responses, 3,529 (70.75%) were correct. The median correct answer rate across all individual questions amounted to 74.26% (Interquartile Range: 32.28, Minimum: 17.64, Maximum: 100). Correct answer rate did not follow a normal distribution (Shapiro-Wilk $p < 0.001$) and was negatively skewed (-0.819).

A unique challenge was encountered while working with GPT-4o: the model occasionally refused to answer certain questions. However, when the same prompt was reattempted or the alternative prompt was used, GPT-4o sometimes provided a response. In one unusual instance, the model generated a sixth answer option (F) and selected it. Since this behavior was inconsistent and was observed only once and only with GPT, it was not systematically analyzed further in this study. While GPT-4o's responses occasionally contained useful information, for the purposes of the present study, any instance in which the model failed to explicitly indicate a valid choice was considered incorrect. By contrast, all other LLMs evaluated provided valid answers in all trials.

Descriptive data regarding the analyzed exams and questions are summarized in Table II. As indicated by the two-proportion z-test results in Table II, all LLMs (individually and in aggregate) were outperformed by candidates. Image resolutions ranged from 48 to 379 dpi (median 152 dpi). Spearman correlation showed no significant association between dpi and the overall best answer rate of any of the LLMs or the correct answer rate of the candidates [GPT: $\rho(106) = -0.147$, $p=0.128$; Grok: $\rho(106) = -0.098$, $p=0.312$; Gemini: $\rho(106) = -0.061$, $p=0.528$; Candidates: $\rho(106) = -0.004$, $p=0.967$].

Cochran's Q test revealed no significant differences among the LLMs (Table III). A comparison of the overall best performance of LLMs to that of candidates revealed no significant difference between candidates and GPT [$z = 1.548$, $p=0.121$]. However, the candidates significantly outperformed Grok [$z = 2.174$, $p=0.030$] and Gemini [$z = 3.215$, $p=0.001$].

A total of 89 questions (82.40%) were answered correctly by at least one LLM in at least one of its runs. In addition, a total of 41 questions (37.96%) were answered correctly by all LLMs in at least one run (see Figure 1 for a detailed distribution of these result)s. As indicated by the results of two proportions z-test, the combined best performance of all LLMs across multiple runs (82.40%) was significantly higher than the candidates' correct answer rate (70.75%) [$z = -2.641$, $p=0.008$].

Furthermore, while repeating the same prompt did not significantly affect the correct answer rates, combining best answers of both runs significantly improved results for GPT (both prompts) and for Grok (HQ-prompt only), but not for Gemini (Table IV). Interestingly, Gemini with the HQ prompt presented its best answers in the first run only to change some of its correct answers to incorrect ones in the second run.

Using different prompts caused a significant difference for GPT and Grok, but not for Gemini. The HQ prompt performed significantly better than the NS prompt. Combining best answers of each prompt significantly increased correct answer rate for all 3 LLMs (Table V).

Next, all LLMs showed significant correlation internally in almost all combinations, with a few coincidental cross correlations with various strength levels. Only GPT answers with the NS prompt showed significant (albeit poor; correlation coefficient <0.3) correlations with the candidate correct answer rate. Full correlation matrix is presented in Supplementary Table I.

As indicated by the results of separate Kruskall-Wallis tests, image types in the questions and content areas did not seem to cause significant difference for candidates; the same trend was revealed by the results of separate FFH tests for overall best performance results of LLMs. For this analysis,

one question containing micrographs was considered in real picture category, while three questions containing DSA images and four questions containing X-ray images were combined together. In addition, as showed by Cochran's Q and McNemar's tests, none of the LLMs significantly outperformed one other in any category. χ^2 statistics are omitted for readability, but the corresponding graphs are presented in Figure 2.

Table II: Summary of Exam Years, Question Counts, and Candidate Versus LLM Performance. The Table Reports the Number of Correct Responses for Each Year Across Candidates and LLMs, Aggregated Over Multiple Runs. (Sum: The sum of four runs per LLM. Total LLM answers: Sum of all 12 runs)

Exam	Candidate	Questions	Correct / Total Candidate Answers (%)	GPT Sum (%)	Grok Sum (%)	Gemini Sum (%)	Correct / Total LLM Answers (%)
2015	38	5	124 / 190 (65.26)	11 / 20 (55.00)	12 / 20 (60.00)	13 / 20 (65.00)	36 / 60 (60.00)
2016	34	10	217 / 340 (63.82)	12 / 40 (30.00)	18 / 40 (45.00)	7 / 40 (17.50)	37 / 120 (30.83)
2017	30	18	343 / 540 (63.52)	36 / 72 (50.00)	44 / 72 (61.11)	39 / 72 (54.17)	119 / 216 (55.09)
2018	41	14	407 / 574 (70.91)	23 / 56 (41.07)	32 / 56 (57.14)	27 / 56 (48.21)	82 / 168 (48.81)
2019	52	12	373 / 624 (59.78)	20 / 48 (41.67)	27 / 48 (56.25)	24 / 48 (50.00)	71 / 144 (49.31)
2022	65	13	645 / 845 (76.33)	29 / 52 (55.77)	24 / 52 (46.15)	28 / 52 (53.85)	81 / 156 (51.92)
2023	62	19	857 / 1178 (72.75)	33 / 76 (43.42)	14 / 76 (18.42)	30 / 76 (39.47)	77 / 228 (33.77)
2024	41	17	563 / 697 (80.77)	31 / 68 (45.59)	36 / 68 (52.94)	44 / 68 (64.71)	111 / 204 (54.41)
Total	363	108	3529 / 4988 (70.75)	195 / 432 (45.14)	207 / 432 (47.92)	212 / 432 (49.07)	614 / 1296 (47.38)
			Two proportions z-test compared to candidates'	z =11.013 **p = <0.001**	z = 9.837 **p = <0.001**	z = 9.346 **p = <0.001**	z = 15.817 **p = < 0.001**

Table III: Correct Answer Rates of LLMs Across Different Prompt Conditions. The Table Reports the Number and Percentage of Correct Responses for Each LLM in Different Test Runs

n = 108	GPT (%)	Grok (%)	Gemini (%)	Cochran's Q
NS 1st Run	48 (44.44)	50 (46.30)	53 (49.07)	$\chi^2(2) = 0.717$, p=0.699
NS 2nd Run	48 (44.44)	51 (47.22)	52 (48.15)	$\chi^2(2) = 0.531$, p=0.767
Best-NS	57 (52.78)	55 (50.93)	55 (50.93)	$\chi^2(2) = 0.160$, p=0.923
HQ 1st Run	45 (41.67)	53 (49.07)	55 (50.93)	$\chi^2(2) = 3.055$, p=0.217
HQ 2nd Run	54 (50.00)	53 (49.07)	52 (48.15)	$\chi^2(2) = 0.120$, p=0.942
Best-HQ	62 (57.41)	61 (56.48)	55 (50.93)	$\chi^2(2) = 1.686$, p=0.430
Best Overall	69 (63.89)	66 (61.11)	61 (56.48)	$\chi^2(2) = 2.042$, p=0.360

NS: *Neurosurgery-prompted responses.* ***HQ:*** *General quiz-prompted responses.* ***Best Overall:*** *The highest accuracy achieved across all four attempts.*

Table IV: Impact of Repeated Questioning on LLM Performance Across Different Prompts. The Table Compares Correct Answer Rates Between the First and Second Attempts, as Well as the Best Combined Performance. Bonferroni Correction (Multiplier of 3) was Applied to Account for Multiple Comparisons. *Indicates Statistically Significant Differences

LLM	Prompt	Cochran's Q	Pair	Correct Answer Rates (%)	McNemar's Exact p-value	Corrected p-value
GPT	NS-Prompt	$\chi^2(2) = 9.000$, $p = 0.011$*	1st run – 2nd run	44.44 - 44.44	1	1
			1st run – Best*	44.44 - 52.78	**0.004***	**0.012***
			2nd run – Best*	44.44 - 52.78	**0.004***	**0.012***
	HQ-Prompt	$\chi^2(2) = 17.360$, $p < 0.001$*	1st run – 2nd run	41.67 - 50.00	0.108	0.323
			1st run – Best*	41.67 - 57.41	**<0.001***	**<0.001***
			2nd run – Best*	50.00 - 57.41	**0.008***	**0.023***
Grok	NS-Prompt	$\chi^2(2) = 4.667$, $p = 0.097$	1st run – 2nd run	46.30 - 47.22	1	1
			1st run – Best	46.30 - 50.93	0.063	0.188
			2nd run – Best	47.22 - 50.93	0.125	0.375
	HQ-Prompt	$\chi^2(2) = 8.000$, $p = 0.018$*	1st run – 2nd run	49.07 - 49.07	1	1
			1st run – Best	49.07 - 56.48	**0.008***	**0.023***
			2nd run – Best	49.07 - 56.48	**0.008***	**0.023***
Gemini	NS-Prompt	$\chi^2(2) = 2.800$, $p = 0.247$	1st run – 2nd run	49.07 - 48.15	1	1
			1st run – Best	49.07 - 50.93	0.500	1
			2nd run – Best	48.15 - 50.93	0.250	0.750
	HQ-Prompt	$\chi^2(2) = 6.000$, $p = 0.050$*	1st run – 2nd run	50.93 - 48.15	0.250	0.750
			1st run – Best	50.93 - 50.93	1	1
			2nd run – Best	48.15 - 50.93	0.250	0.750

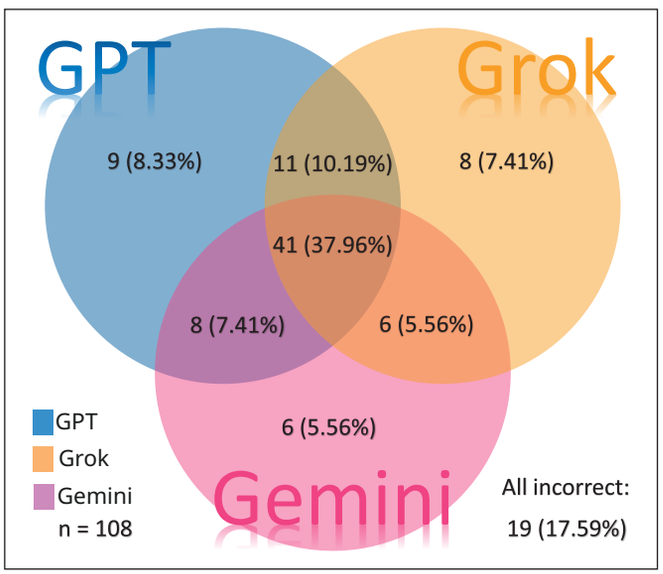

Figure 1: Venn diagram illustrating the number and percentage of questions correctly answered by each of the three LLMs (GPT, Grok, Gemini) in at least one of their four runs. The overlapping areas represent questions correctly answered by multiple LLMs.

DISCUSSION

The present study contributes to the limited yet growing body of research evaluating the performance of LLMs in questions involving visual elements, particularly in medical education and clinical settings. The results not only highlight the potential of these models, but also reveal significant limitations, particularly when compared with previous studies across various specialties.

According to the results, candidates outperformed all LLMs investigated in this study in questions containing visual elements. Repeating the same questions or using different prompts did not significantly alter performance; however, combining the best results from multiple runs led to a significantly improved performance. Only ChatGPT-4o showed a weak correlation with candidate performance when the NS prompt was used. The resolution of the provided images did not affect the performance of the LLMs or the candidates. Performance would be expected to more strongly correlate with dpi in real-world scenarios, where important elements might be subtle. These images in the exams were selected or designed to clearly convey the relevant element(s) to the candidates; therefore, even the lower-resolution images were

Table V: Effect of Different Prompts and Their Combined Results on LLM Performance. The Table Compares Accuracy Rates between NS- and HQ-Prompted Responses, as Well as the Overall Best Combination from Multiple Runs. Bonferroni Correction (Multiplier of 3) was Applied

LLM	Cochran's Q	Pair	Correct Answer Rates (%)	McNemar's exact p-value	Corrected p-value
GPT	$\chi^2(2) = 11.474$, p=0.003*	NS – HQ	52.78 - 57.41	0.359	1
		NS – Ovr*	52.78 - 63.89	**<0.001***	**0.001***
		HQ – Ovr*	57.41 - 63.89	**0.016***	**0.046***
Grok	$\chi^2(2) = 11.375$, p=0.003*	NS – HQ	50.93 - 56.48	0.210	0.630
		NS – Ovr*	50.93 - 61.11	**<0.001***	**0.003***
		HQ – Ovr	56.48 - 61.11	0.063	0.188
Gemini	$\chi^2(2) = 6.000$, p=0.050*	NS – HQ	50.93 - 50.93	1	1
		NS – Ovr	50.93 - 56.48	**0.031***	0.094
		HQ – Ovr	50.93 - 56.48	**0.031***	0.094

NS: *Best answers from two NS-prompted runs.* ***HQ:*** *Best answers from two HQ-prompted runs.* ***Ovr:*** *Best answers from all four runs. *Indicates statistically significant differences.*

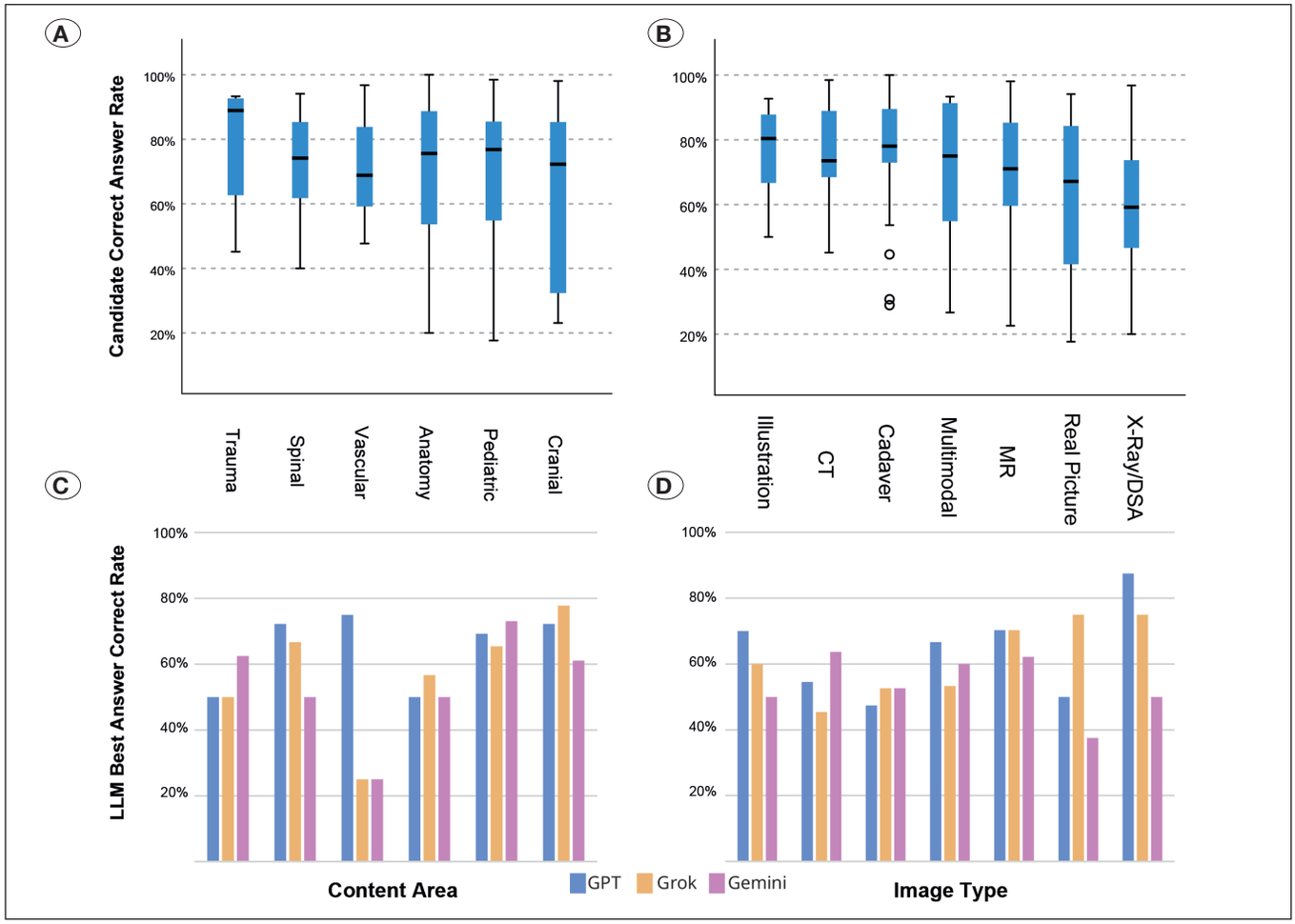

Figure 2: Performance of LLMs and human candidates based on content areas and image types. **A)** Box plot of candidate correct answer rates by content area. **B)** Box plot of candidate correct answer rates by image type. **C)** Bar chart comparing the best performance of each LLM (GPT, Grok, Gemini) by content area. **D)** Bar chart comparing the best performance of each LLM by image type.

sufficient to answer the questions for both the LLMs and the candidates.

Comparison with Other Studies

In a study on the National Medical Licensing Examination in Japan, Liu et al. found that ChatGPT-4o achieved an overall accuracy rate of 89.2%, which dropped to 80.4% for image-based questions (9). By contrast, in the present study, ChatGPT-4o achieved an accuracy rate of 63.89% for image-based questions. This difference may be attributed to the difficulty of the questions and the variety of content that LLMs encountered in different exams. Furthermore, consistently with our finding, Liu et al. also reported that Gemini 1.5 Pro performed worse than both GPT-4 and ChatGPT-4o on image-based questions (74.6%) (9).

However, our analysis further revealed that Gemini's responses to image-based questions were inconsistent, as it occasionally changed correct answers upon reattempts. Unlike Liu et al.'s study, which presented questions in a direct exam format without specialized prompts, we employed two distinct prompting strategies (the NS prompt and the HQ prompt). While our results indicated that prompting strategies influence performance, they did not provide a substantial improvement, particularly for image-based questions.

Furthermore, Lin et al. reported that ChatGPT-4o reported 77.19% accuracy on image-based questions, whereas Gemini performed significantly worse on image-based questions with only 42.11% accuracy (8). In line with this evidence, we found that Gemini demonstrated low performance in questions based on complex visual content. However, the success of ChatGPT-4o in Lin et al.'s study was quite high as compared to our results (8).

In contrast to Liu et al. and Lin et al., Gill et al. found ChatGPT-4o to have 39.58% accuracy in image-based questions, while Gemini Advanced remained at 33.33%. The authors reported that ChatGPT-4o performed significantly better than Gemini Advanced ($p<0.013$) (5,8,9). In our findings, both models were more successful, and no statistically significant difference was observed between the models.

In yet another relevant study, Takagi et al. (18) reported that ChatGPT-4v achieved 71.9% accuracy for image-based questions on the Japan National Medical Licensing Examination, which was below the average performance of human candidates (85.0%) ($p<0.001$). For the questions with tables, the accuracy rate amounted to mere 35%, which was significantly lower than human performance (83.9%) ($p<0.001$) (18). While such significant difference was evident in our comparison of human performance vs. that of ChatGPT-4o, candidates were statistically significantly more successful than Grok and Gemini.

Furthermore, Nguyen et al. reported that ChatGPT-4 achieved between 36% and 50% accuracy on image-based questions (11). The authors showed that encouragement and responsibility disclaimer prompts significantly improved ChatGPT-4's performance on image-based questions (up to 65% increase). By contrast, when medico-legal or patient care prompts were used, the model more frequently avoided answering the questions and its performance deteriorate. Our study complements the findings of Nguyen et al. by evaluating the effect of specific prompts for neurosurgical scenarios (11). Specifically, while the use of prompts with neurosurgical roles (e.g., "Think of yourself as a Brain Surgeon") improved performance on image-based questions, this improvement was not sufficient to increase overall success rates. This suggests that prompt engineering alone may not be sufficient to improve success on image-based questions.

In another relevant study evaluating the performance of ChatGPT-4 in the American Registry of Radiologic Technologists certification examination, the model tested by Al-Naser et al.'s achieved an accuracy rate of 45.6% on image-based questions (1). Similarly, in the study by Sawamura et al., ChatGPT-4o attained an accuracy rate of 35.4% on questions containing visuals and tables in the Japan National Physiotherapist Examination (17). Yet another study by Noda et al. reported that GPT-4v performed with an accuracy rate of 41.3% on image-based questions in the 2023 Japan Otolaryngology Specialist Examination (12). Importantly, this study highlighted that English translations and additional prompts significantly improved the model's accuracy on image-based questions, thereby enhancing its overall performance. These results were lower compared to the 68% accuracy rate reported by Nakao et al. for GPT-4v on image-based questions in the Japan National Medical Licensing Examination. Overall, the results of the present study more closely aligned with the findings reported by Nakao et al (10).

Since success may vary depending on the exam names, questions, and language models, several important points should be made. By now, the failure observed in the early days in text-based questions has almost completely disappeared. If there is a failure only in text-based questions, the question stem should be analyzed as in the results of our previous study. However, when the results reported in the literature and our results in image-based questions are compared with only text-based questions, it becomes evident that sufficient success has not yet been achieved.

Unique Contributions of the Present Study

While previous studies have predominantly examined LLM performance in general medical and specialty-specific examinations, the present study offers unique insights into the neurosurgical domain, where image-based questions are particularly complex and critical. Compared to prior research, our findings further underscore the variability in LLM performance based on question complexity, domain specificity, and prompt design. For instance, in our data, the combination of multiple runs and prompts significantly improved the overall correct answer rates for ChatGPT-4o, Grok, and Gemini—a methodological aspect frequently overlooked in prior evaluations.

Furthermore, in contrast to previous studies in that primarily analyzed static visual data in nephrology and otolaryngology (e.g., biopsy slides, clinical photographs) (3,16), the Turkish Neurosurgical Society Proficiency Board Exams that incorporates visual elements such as imaging studies and operative

diagrams present a uniquely demanding test environment. The dynamic and context-dependent nature of neurosurgical visuals may explain the lower performance of LLMs in this domain compared to others.

To the best of our knowledge, this study is the first to evaluate Grok's performance on image-based questions. Developed by xAI, Grok was introduced as a language model with real-time knowledge integration capabilities, primarily analyzed in the context of general LLM performance and social media applications. However, this study is the first to scientifically evaluate Grok in a medical context, specifically within the highly specialized field of neurosurgery. In addition, the present study is the first investigation of multimodal AI models, including Grok, on visual-dominant assessments such as the Turkish Neurosurgical Society Proficiency Board Exams. This pioneering contribution not only expands the understanding of Grok's capabilities, but also highlights the transformative potential of AI technologies in medical education and standardized examinations. In so doing, this study marks a significant milestone in AI-related neurosurgical research.

Limitations and Implications for Future Research

The consistency of findings across studies highlights the need for targeted improvements in LLM capabilities. While the introduction of multimodal inputs in models like ChatGPT-4o represents a significant advancement, the relatively low accuracy rates for image-based questions suggest that current architectures are not yet optimized for integrating visual and textual data effectively. In this study, we observed each model in its native state, without fine-tuning, pre-training, or contextual enrichment. While pre-training can enhance LLMs' image interpretation capabilities (21), contextual enrichment may be less effective, at least for neurosurgical topics (7). Incorporating contextual enrichment with images could significantly boost LLM performance, with some models potentially demonstrating greater adaptability than others after such augmentation. Accordingly, future research should explore these approaches to facilitate reliable AI-based decision support systems for clinical use.

Another limitation of the present study is that questions were only presented in their original language (Turkish). None of the investigated LLMs explicitly discloses whether it is language-aware or language-agnostic. However, it is highly likely that these models performed better in languages with more available resources in their training datasets. The present literature only investigated Turkish performance of GPT for social sciences, yielding conflicting results (15,19). Many fields—including psychology, sociology, communications, political science, and computer science—use computational methods to analyze text data. However, existing text analysis methods have a number of shortcomings. Dictionary methods, while easy to use, are often not very accurate when compared to recent methods. Machine learning models, while more accurate, can be difficult to train and use. We demonstrate that the large-language model GPT is capable of accurately detecting various psychological constructs (as judged by manual annotators. Accordingly, dedicated studies would be needed to evaluate the neurosurgical competency of LLMs in different languages and to determine whether linguistic factors influence model performance.

CONCLUSION

While numerous previous studies demonstrated remarkable capabilities of LLMs in answering neurosurgical or medical exam questions, these models remain inadequate when confronted with questions containing visual elements. Prompt engineering influences model performance in real-world scenarios; however, human candidate performance can be surpassed only by combining the best responses from multiple runs across multiple LLMs.

The results of the present study highlight both the potential and limitations of multimodal AI models in high-stakes medical assessments. While their capabilities continue to advance, their performance in image-based scenarios highlights critical areas requiring further improvement. Taken together, our findings emphasize the need for collaboration between AI developers and medical professionals to enhance model training, particularly in integrating complex, real-world clinical data. Such advancements could revolutionize not only medical education, but also clinical decision making, thereby ultimately enhancing patient outcomes.

ACKNOWLEDGMENTS

The authors thank Prof. Hakan Emmez and Prof. Ender Koktekir for their support at every step of this research. We also acknowledge the Turkish Neurosurgical Society Board of Directors for providing access to the exam dataset.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials: Complete exams are available at https://turknorosirurji.org.tr/TNYK/ (in Turkish). Additional data are available upon reasonable request.

Disclosure: The authors declare no competing interests.

Ethics Statement: This study evaluated questions containing visual elements from the last eight written Turkish Neurosurgical Society Proficiency Board Exams. The exam consisted of multiple-choice questions, each with 5 answer options. The questions and answer keys are publicly available on the Turkish Neurosurgical Society website. Candidate responses for all questions were obtained in an anonymized format, upon obtaining the permission of the Turkish Neurosurgery Association Board of Directors. No additional approval was required for this investigation by local and regional regulations.

AUTHORSHIP CONTRIBUTION

Study conception and design: AS, MCS
Data collection: GE, OYT
Analysis and interpretation of results: AS, BS, MBS
Draft manuscript preparation: AS, MCS
Critical revision of the article: GE, OYT, BS, MBS
All authors (AS, GE, OYT, BS, MBS, MCS) reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Al-Naser Y, Halka F, Ng B, Mountford D, Sharma S, Niure K, Yong-Hing C, Khosa F, Van der Pol C: Evaluating artificial intelligence competency in education: Performance of Chatgpt-4 in the american registry of radiologic technologists (ARRT) radiography certification exam. Acad Radiol 32:597-603, 2025. https://doi.org/10.1016/j.acra.2024.08.009
2. base64: base64 - Base16, Base32, Base64, Base85 Data Encodings [Internet]. Python documentation, 2024. Available from: https://docs.python.org/3/library/base64.html
3. Bulduk EB, Yilmaz C: Turkish board of neurological surgery. Turk Neurosurg 29:121-126, 2019. https://doi.org/10.5137/1019-5149.JTN.22521-18.1
4. Chan YH: Biostatistics 104: Correlational analysis. Singapore Med J 44:614-619, 2003
5. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S: Comparison of gemini advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. Cureus 16: e69612, 2024. https://doi.org/10.7759/cureus.69612
6. Google LLC: Introducing Gemini: Our largest and most capable AI model [Internet]. Google 2023. Available from: https://blog.google/technology/ai/google-gemini-ai/
7. Guo E, Gupta M, Sinha S, Rössler K, Tatagiba M, Akagami R, Al-Mefty O, Sugiyama T, Stieg PE, Pickett GE, Lotbiniere-Bassett M de, Singh R, Lama S, Sutherland GR: NeuroGPT-X: Toward a clinic-ready large language model. J Neurosurg 140:1041-1053, 2023. https://doi.org/10.3171/2023.7.JNS23573
8. Lin SY, Hsu YY, Ju SW, Yeh PC, Hsu WH, Kao CH: Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023. Digit Health 10:20552076241291404, 2024. https://doi.org/10.1177/20552076241291404
9. Liu M, Okuhara T, Dai Z, Huang W, Gu L, Okada H, Furukawa E, Kiuchi T: Evaluating the Effectiveness of advanced large language models in medical knowledge: A comparative study using Japanese national medical examination. Int J Med Inform 193:105673, 2025. https://doi.org/10.1016/j.ijmedinf.2024.105673
10. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, Yoshikawa T, Abe O: Capability of GPT-4V(ision) in the Japanese national medical licensing examination: Evaluation study. JMIR Med Educ 10:e54393, 2024. https://doi.org/10.2196/54393
11. Nguyen D, MacKenzie A, Kim YH: Encouragement vs. liability: How prompt engineering influences ChatGPT-4's radiology exam performance. Clin Imaging 115:110276, 2024. https://doi.org/10.1016/j.clinimag.2024.110276
12. Noda M, Ueno T, Koshu R, Takaso Y, Shimada MD, Saito C, Sugimoto H, Fushiki H, Ito M, Nomura A, Yoshizaki T: Performance of GPT-4V in answering the Japanese otolaryngology board certification examination questions: Evaluation study. JMIR Med Educ 10:e57054, 2024. https://doi.org/10.2196/57054
13. OpenAI: Hello GPT-4o [Internet] Available from: https://openai.com/index/hello-gpt-4o/
14. OpenAPI Initiative: OAI/learn.openapis.org [Internet], 2024. Available from: https://github.com/OAI/learn.openapis.org
15. Rathje S, Mirea D-M, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ: GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci U S A 121:e2308950121, 2024. https://doi.org/10.1073/pnas.2308950121
16. Sahin MC, Sozer A, Kuzucu P, Turkmen T, Sahin MB, Sozer E, Tufek OY, Nernekli K, Emmez H, Celtikci E: Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med 169:107807, 2024. https://doi.org/10.1016/j.compbiomed.2023.107807
17. Sawamura S, Kohiyama K, Takenaka T, Sera T, Inoue T, Nagai T: Performance of ChatGPT 4.0 on Japan's national physical therapist examination: A comprehensive analysis of text and visual question handling. Cureus 16:e67347, 2024. https://doi.org/10.7759/cureus.67347
18. Takagi S, Koda M, Watari T: The Performance of ChatGPT-4V in interpreting images and tables in the Japanese medical licensing exam. JMIR Med Educ 10:e54283, 2024. https://doi.org/10.2196/54283
19. Unlutabak B, Bal O: Theory of mind performance of large language models: A comparative analysis of Turkish and English. Computer Speech Language 89:101698, 2025. https://doi.org/10.1016/j.csl.2024.101698
20. xAI: Welcome | xAI [Internet] Available from: https://x.ai/
21. Yu S, Sun W, Mi D, Jin S, Wu X, Xin B, Zhang H, Wang Y, Sun X, He X: Artificial intelligence diagnosing of oral lichen planus: A comparative study. Bioengineering (Basel) 11:1159, 2024. https://doi.org/10.3390/bioengineering11111159

Supplementary Table I: Complete Correlation Matrix of the Study Variables

	Candidate Correct Answer Rate	GPT NS-Prompt 1st Run	GPT NS-Prompt 2nd Run	GPT NS-Prompt Best
Candidate Correct Answer Rate		ρ(106) = 0.267, p = 0.005	ρ(106) = 0.234, p = 0.015	ρ(106) = 0.267, p = 0.005
GPT NS-Prompt 1st Run	ρ(106) = 0.267, p = 0.005		ρ(106) = 0.663, p < 0.001	ρ(106) = 0.846, p < 0.001
GPT NS-Prompt 2nd Run	ρ(106) = 0.234, p = 0.015	ρ(106) = 0.663, p < 0.001		ρ(106) = 0.846, p < 0.001
GPT NS-Prompt Best	ρ(106) = 0.267, p = 0.005	ρ(106) = 0.846, p < 0.001	ρ(106) = 0.846, p < 0.001	
GPT HQ-Prompt 1st Run	ρ(106) = 0.168, p = 0.081	ρ(106) = 0.491, p < 0.001	ρ(106) = 0.567, p < 0.001	ρ(106) = 0.536, p < 0.001
GPT HQ-Prompt 2nd Run	ρ(106) = 0.188, p = 0.051	ρ(106) = 0.634, p < 0.001	ρ(106) = 0.708, p < 0.001	ρ(106) = 0.686, p < 0.001
GPT HQ-Prompt Best	ρ(106) = 0.156, p = 0.106	ρ(106) = 0.544, p < 0.001	ρ(106) = 0.695, p < 0.001	ρ(106) = 0.648, p < 0.001
Grok NS-Prompt 1st Run	ρ(106) = 0.101, p = 0.297	ρ(106) = 0.552, p < 0.001	ρ(106) = 0.440, p < 0.001	ρ(106) = 0.506, p < 0.001
Grok NS-Prompt 2nd Run	ρ(106) = 0.168, p = 0.083	ρ(106) = 0.498, p < 0.001	ρ(106) = 0.423, p < 0.001	ρ(106) = 0.486, p < 0.001
Grok NS-Prompt Best	ρ(106) = 0.121, p = 0.213	ρ(106) = 0.505, p < 0.001	ρ(106) = 0.431, p < 0.001	ρ(106) = 0.518, p < 0.001
Grok HQ-Prompt 1st Run	ρ(106) = 0.087, p = 0.370	ρ(106) = 0.427, p < 0.001	ρ(106) = 0.277, p = 0.004	ρ(106) = 0.335, p < 0.001
Grok HQ-Prompt 2nd Run	ρ(106) = 0.136, p = 0.159	ρ(106) = 0.501, p < 0.001	ρ(106) = 0.352, p < 0.001	ρ(106) = 0.409, p < 0.001
Grok HQ-Prompt Best	ρ(106) = 0.095, p = 0.330	ρ(106) = 0.447, p < 0.001	ρ(106) = 0.297, p = 0.002	ρ(106) = 0.367, p < 0.001
Gemini NS-Prompt 1st Run	ρ(106) = 0.111, p = 0.254	ρ(106) = 0.240, p = 0.012	ρ(106) = 0.352, p < 0.001	ρ(106) = 0.335, p < 0.001
Gemini NS-Prompt 2nd Run	ρ(106) = 0.138, p = 0.156	ρ(106) = 0.331, p < 0.001	ρ(106) = 0.406, p < 0.001	ρ(106) = 0.392, p < 0.001
Gemini NS-Prompt Best	ρ(106) = 0.139, p = 0.152	ρ(106) = 0.282, p = 0.003	ρ(106) = 0.393, p < 0.001	ρ(106) = 0.370, p < 0.001
Gemini HQ-Prompt 1st Run	ρ(106) = 0.086, p = 0.377	ρ(106) = 0.282, p = 0.003	ρ(106) = 0.319, p = 0.001	ρ(106) = 0.296, p = 0.002
Gemini HQ-Prompt 2nd Run	ρ(106) = 0.076, p = 0.434	ρ(106) = 0.294, p = 0.002	ρ(106) = 0.331, p < 0.001	ρ(106) = 0.318, p = 0.001
Gemini HQ-Prompt Best	ρ(106) = 0.086, p = 0.377	ρ(106) = 0.282, p = 0.003	ρ(106) = 0.319, p = 0.001	ρ(106) = 0.296, p = 0.002
GPT Best of 4	ρ(106) = 0.188, p = 0.052	ρ(106) = 0.672, p < 0.001	ρ(106) = 0.672, p < 0.001	ρ(106) = 0.795, p < 0.001
Grok Best of 4	ρ(106) = 0.104, p = 0.283	ρ(106) = 0.408, p < 0.001	ρ(106) = 0.369, p < 0.001	ρ(106) = 0.425, p < 0.001
Gemini Best of 4	ρ(106) = 0.101, p = 0.300	ρ(106) = 0.259, p = 0.007	ρ(106) = 0.372, p < 0.001	ρ(106) = 0.329, p < 0.001

Code 1: Prompts in their code form that was used for the purpose of this study. Please note, lines starting with # were used alternatingly to create all combinations of LLMs and prompts twice. The 'base_url' variable (not shown here) was modified as necessary.

```
# model="gpt-4o",

# model="grok-vision-beta",

# model="gemini-1.5-pro",

messages=[

    # {"role": "system", "content": "You are a neurosurgeon presented with a multiple-choice board exam question that includes an image. Carefully analyze the question and select the most appropriate answer from the provided choices. Format your response as follows: 'Answer choice - Your reasoning.' For example: 'A - Your reasoning.' Begin with the answer choice, followed by ' - ', and then provide your reasoning."},

    # {"role": "system", "content": "Help me with my quiz tomorrow! Carefully analyze the question and select the most appropriate answer from the provided choices. Format your response as follows: 'Answer choice - Your reasoning.' For example: 'A - Your reasoning.' Begin with the answer choice, followed by ' - ', and then provide your reasoning."},

    {
      "role": "user",
      "content": [
        {"type": "text", "text": question_text},
        {
            "type": "image_url",
            "image_url": {
                "url": f"{base64_image}",
          },
        },
      ],
    }
],
```