

Article

Artificial Intelligence in Plastic Surgery Education: Insights from Parallel Turkish and English Versions of a Board Examination

Ibrahim Güler ^{1,2,*} , Armin Kraus ¹ , Gerrit Grieb ^{3,4} , Uzay Cambaz ⁵, Henrik Stelling ^{2,6,†} and Cenk Demirdover ^{7,8,9,†} 

¹ Department of Plastic, Aesthetic and Hand Surgery, Otto-von-Guericke University, 39120 Magdeburg, Germany; armin.kraus@med.ovgu.de

² Department of Health Management, Friedrich-Alexander-Universität Erlangen-Nürnberg, 90403 Nürnberg, Germany; henrik.stelling@fau.de

³ Department of Plastic Surgery and Hand Surgery, Gemeinschaftskrankenhaus Havelhoehe, 14089 Berlin, Germany; gerritgrieb@gmx.de

⁴ Department of Plastic Surgery and Hand Surgery, Medical Faculty, Rheinisch-Westfälische Technische Hochschule Aachen University, 52074 Aachen, Germany

⁵ Faculty of Medicine, Eberhard Karls University of Tübingen, 72076 Tübingen, Germany; uzay.cambaz@student.uni-tuebingen.de

⁶ Practices for Nuclear Medicine, 12157 Berlin, Germany

⁷ Turkish Society of Plastic, Reconstructive and Aesthetic Surgeons (TSPRAS), 06530 Ankara, Turkey; cenkddr@gmail.com

⁸ European Board of Plastic, Reconstructive and Aesthetic Surgery (EBOPRAS), 9700 RB Groningen, The Netherlands

⁹ European Society of Plastic, Reconstructive and Aesthetic Surgery (ESPRAS), 5001 Aarau, Switzerland

* Correspondence: ibrahim.gueler@med.ovgu.de

† These authors contributed equally to this work.

Featured Application

For assessing specialized factual knowledge in demanding disciplines such as medicine, here represented by plastic surgery, large language models (LLMs) can serve as multiple-choice question (MCQ) study aids. Known human performance gaps between a first (L1) and a second (L2) language were not reflected in the models' accuracy across the Turkish and English question versions, so these tools can also be used for board preparation in languages other than English. Free model tiers performed comparably to paid ones, making these tools accessible in resource-limited settings. Their outputs should be critically appraised and used under specialist supervision rather than treated as a substitute for clinical competence.

Abstract

Background: Large language models (LLMs) increasingly pass medical board examinations and aid clinical knowledge retrieval and decision support; validating their specialist knowledge is a prerequisite for safe use. Two limitations weaken existing evidence: many studies reuse public questions that may be in training data, and most evaluations use only English-language examinations. Plastic surgery serves here as a representative, highly specialized subfield of medicine. **Methods:** We used the non-public board examination of the Turkish Society of Plastic, Reconstructive, and Aesthetic Surgeons (TSPRAS): 100 single-best-answer (SBA) items in official parallel Turkish and English versions. Six contemporary LLMs from three developers, in matched free and paid tiers, completed six runs per language (7200 responses) under consistent English instruction. Accuracy, inter-run reliability, and cross-language and tier differences were analyzed. **Results:** All models exceeded the 60% passing threshold (71.5–86.4% aggregate), with almost perfect inter-run agreement (Cohen's $\kappa \geq 0.843$). Question language produced no significant difference after correction,



Academic Editor: Thomas Lindner

Received: 9 July 2026

Revised: 22 July 2026

Accepted: 23 July 2026

Published: 30 July 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](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and item difficulty correlated strongly across languages (Pearson $r = 0.928$); divergent items reflected content, not language. Paid variants offered only incremental advantages. Conclusions: Current LLMs showed no significant accuracy difference between the Turkish and English examination versions under a constant English prompt, supporting supervised study-aid use on this MCQ format.

Keywords: artificial intelligence; benchmarking; large language models; generative artificial intelligence; plastic and reconstructive surgery; cross-language evaluation; medical education; board examination; multilingual assessment; reproducibility

1. Introduction

Artificial intelligence (AI), and more specifically, generative artificial intelligence (GenAI), is increasingly shaping everyday life, medicine, and medical education. While many current clinical applications remain highly specialized artificial narrow intelligence (ANI) systems designed for specific tasks, such as automated fracture detection or image interpretation, general-purpose AI (GPAI) tools have recently become accessible to millions of users worldwide. Among these, large language models (LLMs) and, more recently, multimodal large language models (MLLMs) capable of processing information beyond text alone have rapidly entered everyday medical practice. Their use extends from educational applications, including examination preparation and knowledge revision, to informal clinical support tasks such as information retrieval, drafting, and decision support. As a result, these systems are also being adopted by plastic surgery trainees and specialists, making their capabilities, limitations, and educational value an important area of research. Within plastic, reconstructive, and aesthetic surgery, AI is increasingly being explored for preoperative planning, outcome prediction, and clinical decision support, and the safe adoption of general-purpose LLMs in this field depends first on establishing how reliably they command specialist knowledge [1–5].

Contemporary LLMs have repeatedly demonstrated passing-level performance on multiple-choice question (MCQ) examinations in medicine, including both general medical licensing examinations and specialty-specific assessments [6–10]. Similar findings have also been reported in plastic surgery, where LLMs have achieved remarkable performance on the American Society of Plastic Surgeons (ASPS) In-Service Examination, also known as the Plastic Surgery In-Training Examination (PSITE), with performance generally improving across successive model generations and over time [11–19]. Although passing-level performance on medical and specialty-specific MCQ examinations has been demonstrated repeatedly, several aspects remain open on both the medical and the technical side.

First, most studies have used examination questions that are accessible online, whether through official release or informal circulation, and that may therefore already be present in model training data. Where this is the case, high scores may reflect memorization of previously encountered items rather than genuine domain knowledge, which substantially weakens the validity of the benchmark.

Second, the evidence base specific to plastic surgery is narrow. Beyond the PSITE, few examinations have been studied, with only isolated reports from other settings such as the Taiwan plastic surgery board examination. Examinations from other countries and national curricula therefore remain largely unrepresented [20,21].

Third, from a technical perspective, evaluations are conducted overwhelmingly in English, even though many trainees worldwide sit their board examinations in another language. In human examinees, performance declines when testing occurs not in the first

(L1) but in a second language (L2), raising the question of whether LLM performance is similarly language-dependent [22–25]. Whether examination language influences LLM performance has received little attention in plastic surgery, and to our knowledge, no prior study has evaluated LLMs on a national plastic surgery board examination using non-public items administered in parallel language versions.

Fourth, while free and paid model tiers have been compared, whether paid subscriptions actually deliver superior specialist knowledge is rarely tested, and barely at all in plastic surgery, despite trainees and departments routinely paying for premium access.

The National Board Examination of the Turkish Society of Plastic, Reconstructive, and Aesthetic Surgeons (TSPRAS) provides an appropriate vehicle to address these gaps. The theoretical examination comprises 100 single-best-answer MCQs covering the national core curriculum and was provided by the TSPRAS in two official parallel versions, Turkish and English, representing content-equivalent forms of the same high-stakes specialist assessment [26,27]. This permits a controlled within-instrument evaluation of cross-lingual performance, and we used it to address the questions raised above in a single bilingual setting.

2. Materials and Methods

2.1. Study Design

This study used a repeated-measures benchmarking design to evaluate the examination performance and response consistency of six LLMs on a standardized national specialist board examination in plastic and reconstructive surgery. Each administration was treated as an independent measurement occasion, and model outputs were analyzed at the level of observed examination performance. The study did not involve human participants, biological material, or personal data. Ethical review and approval were therefore not required.

2.2. Assessment Instrument

The assessment instrument was the original theoretical component of the TSPRAS exam (2017 administration), provided by the TSPRAS in both its official Turkish and official English versions. The exam comprises 100 single-best-answer MCQs, each with five response options and one correct answer, covering the full national core curriculum of plastic surgery, including reconstructive, aesthetic, hand, microsurgical, and craniofacial surgery. The pass mark for the theoretical component is 60% of items answered correctly, an absolute standard fixed by the Society's examination regulations that is not norm-referenced to cohort performance [26,27].

The TSPRAS certification process consists of this written examination and a subsequent practical clinical component in the form of an Objective Structured Clinical Examination (OSCE). The present study focused exclusively on the theoretical 100-item component, as only this section permits standardized, repeated, and language-parallel administration to LLMs under controlled conditions. The two language versions are content-equivalent translations of the same items, enabling a direct within-examination comparison of performance across languages [26,27]. Both language versions and the official answer key were screened prior to analysis by the board-certified plastic surgeons among the authors; no item or keyed answer was found to be erroneous.

2.3. Models Evaluated

Six proprietary LLMs were evaluated, stratified by developer and subscription tier. Paid variants comprised SuperGrok 4.1 (xAI, San Francisco, CA, USA), ChatGPT-5.2 Pro (OpenAI, San Francisco, CA, USA), and Gemini 3 Pro (Google LLC, Mountain View, CA, USA). Free variants comprised Grok 4.1 (xAI), ChatGPT-5.2 (OpenAI), and Gemini 3

(Google LLC). This design of three matched developer dyads, each comprising a free and a paid variant from the same developer, enabled comparison of the free and paid products offered by each developer. All models were evaluated in their publicly available versions as accessible in February 2026 and were not fine-tuned on TSPRAS materials.

2.4. Testing Protocol

Each model was administered both language versions of the 100-item examination across six independent sessions, yielding 12 complete examinations per model and 72 examination runs in total (7200 individual item responses). Sessions were separated by a minimum of 24 h and initiated from fresh, context-free conversation threads to preclude cross-session memory contamination. All models were accessed exclusively through their official public web interfaces, without use of application programming interfaces (APIs), reflecting real-world user access. A standardized zero-shot prompt was used for all models, runs, and both languages, instructing the model to answer board-level MCQs and to output only the selected letter (A–E) without explanation. All prompt instructions were provided in English for both the Turkish and English question conditions, thereby holding the instruction (prompt) language constant in order to isolate the effect of the content (item) language, that is, the language in which the examination items themselves were presented. No retrieval-augmented generation (RAG), external tools, memory, or fine-tuning were used. This format followed established practices for structured LLM evaluation in medical education [6,7,28].

2.5. Statistical Analysis

Item-level accuracy was scored dichotomously and aggregated to the percentage of correctly answered items per run; mean accuracy and standard deviation (SD) were computed across the six runs for each model and language condition. Ninety-five percent confidence intervals (CIs) for accuracy were estimated by an item-clustered bootstrap that resampled the 100 items with replacement (20,000 resamples), retaining all repeated responses within each resampled item so that the dependence among repeated answers to the same item is preserved. Inter-run reliability was quantified using Cohen's κ computed on the full categorical responses (A–E), calculated pairwise for all 15 unique run pairs within each model and language condition and reported as the mean pairwise value; values were interpreted according to the Landis and Koch classification [29]. For paired comparisons, modal responses were derived by majority vote across the six runs (≥ 4 of 6); items without a clear majority were excluded from these paired comparisons only, whereas run-level accuracy, inter-run reliability, and the item-level difficulty analysis used all responses.

Subscription-tier and cross-language differences were tested using McNemar's exact test on matched modal responses. To assess tier effects at the run level, per-run accuracy was additionally compared between matched free and paid variants using Welch's *t*-test. Item difficulty (proportion correct per item) was compared across languages using Pearson and Spearman correlation, with a bootstrap 95% CI for Pearson *r* (5000 resamples) and per-model Cohen's κ on item-level modal responses. All *p*-values were adjusted for multiple comparisons using the Holm–Bonferroni procedure across the full family of tests. Analyses were performed using Python 3.12 (Python Software Foundation, Wilmington, DE, USA) with standard statistical libraries.

3. Results

3.1. Overall Performance

Across all 72 examination runs and both language versions, individual run accuracy ranged from 52.0% (ChatGPT-5.2, Turkish, run 5) to 90.0% (Gemini 3 Pro, Turkish, run 3).

Aggregated across languages and runs, mean performance ranged from 71.5% (ChatGPT-5.2) to 86.4% (Gemini 3 Pro), with all six models exceeding the 60% specialist passing threshold in the large majority of administrations (Table 1, Figure 1). The highest aggregated accuracy was observed for the leading paid model (86.4%), followed closely by several free and paid variants clustered between 79% and 83%.

Table 1. Mean accuracy ($\pm$ SD) and 95% confidence intervals by model, subscription tier, and language version across six independent runs per condition (100 items per run).

Model	Tier	English Mean $\pm$ SD (%)	Turkish Mean $\pm$ SD (%)	Aggregated Mean $\pm$ SD (%)	Aggregated 95% CI
Gemini 3	Free	80.7 $\pm$ 1.0	85.2 $\pm$ 1.3	82.9 $\pm$ 2.6	76.0–89.2
ChatGPT-5.2	Free	74.0 $\pm$ 1.1	69.0 $\pm$ 8.4	71.5 $\pm$ 6.3	63.6–79.1
Grok 4.1	Free	82.0 $\pm$ 1.4	81.8 $\pm$ 1.7	81.9 $\pm$ 1.5	75.4–88.0
Gemini 3 Pro	Paid	86.2 $\pm$ 0.8	86.7 $\pm$ 1.9	86.4 $\pm$ 1.4	80.3–91.9
ChatGPT-5.2 Pro	Paid	79.8 $\pm$ 0.4	78.5 $\pm$ 1.8	79.2 $\pm$ 1.4	71.8–86.1
SuperGrok 4.1	Paid	81.3 $\pm$ 2.7	82.7 $\pm$ 2.2	82.0 $\pm$ 2.4	75.4–88.2

Aggregated values pool both language versions ($n = 1200$ item responses per model). Confidence intervals are 95% item-clustered bootstrap intervals. SD, standard deviation; CI, confidence interval.

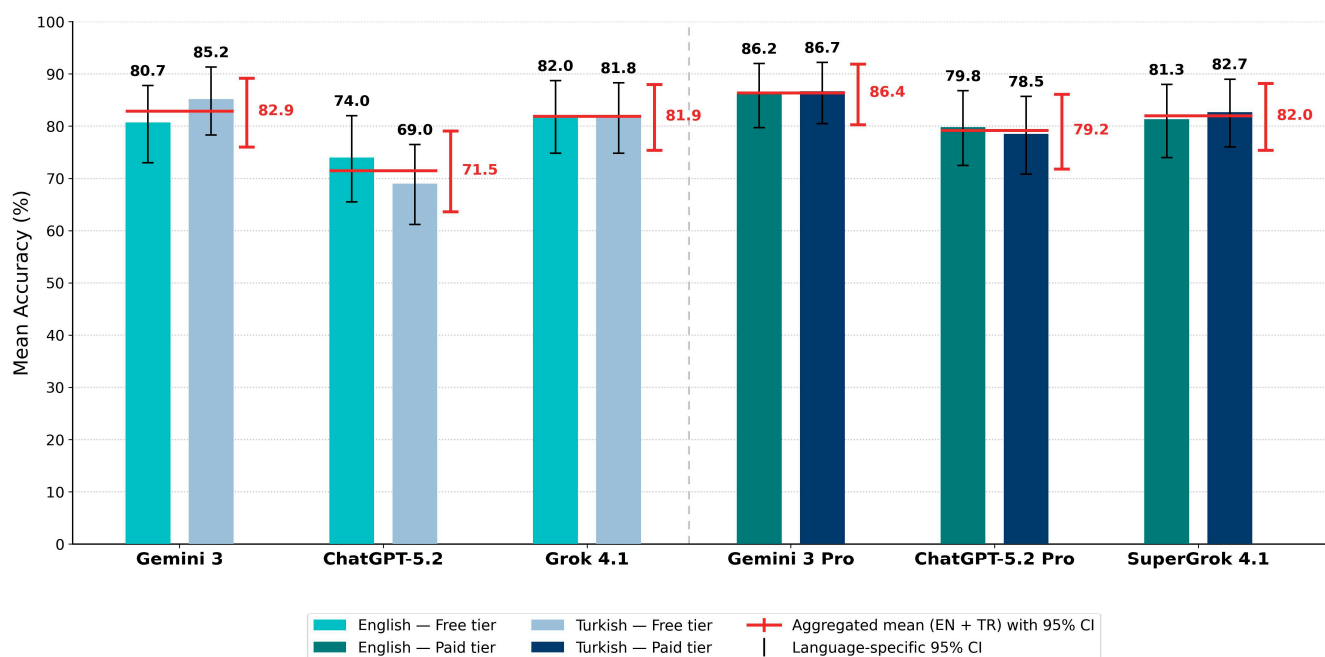


Figure 1. Mean accuracy of six LLMs on the theoretical component of the TSPRAS examination, by language version and subscription tier across six independent runs per condition (100 items per run). Color hue encodes language version, and color depth encodes subscription tier, as shown in the legend. Red horizontal bars indicate aggregated mean accuracy across both language conditions; right-side vertical brackets indicate aggregated 95% item-clustered bootstrap confidence intervals. Error bars on individual bars represent language-specific 95% item-clustered bootstrap CIs. The dashed vertical line separates free-tier (left) from paid-tier (right) models. CI, confidence interval.

3.2. Inter-Run Reliability

Mean pairwise Cohen's κ was high across virtually all conditions, with every model yielding an aggregated value above 0.80, indicating almost perfect agreement [29] (Table 2). The single condition-level exception was ChatGPT-5.2 in the Turkish condition ($\kappa = 0.759$, substantial agreement [29]), attributable to a single atypical run (52.0% accuracy, approximately 20 percentage points below that model's typical Turkish performance) rather than

to systematic response instability; this run was retained in all analyses, and the model's aggregated κ remained within the “almost perfect” [29] range (0.843).

Table 2. Inter-run reliability expressed as mean pairwise Cohen's κ across all 15 run pairs per model and language condition.

Model	Tier	κ (English)	κ (Turkish)	κ (Aggregated)	Interpretation (Aggregated)
Gemini 3	Free	0.942	0.936	0.939	Almost perfect
ChatGPT-5.2	Free	0.927	0.759	0.843	Almost perfect
Grok 4.1	Free	0.911	0.884	0.897	Almost perfect
Gemini 3 Pro	Paid	0.930	0.911	0.920	Almost perfect
ChatGPT-5.2 Pro	Paid	0.902	0.917	0.909	Almost perfect
SuperGrok 4.1	Paid	0.892	0.885	0.888	Almost perfect

Computed on the full categorical responses (A–E); κ interpreted per Landis and Koch [29]: >0.80, almost perfect; 0.61–0.80, substantial. κ , Cohen's kappa.

3.3. Cross-Language Consistency

McNemar's exact tests comparing English and Turkish modal responses within each model yielded no statistically significant differences after Holm–Bonferroni correction (all corrected comparisons non-significant; Table 3). In one model (Gemini 3), all six discordant items favored the Turkish version (6 versus 0 of 98 retained items; uncorrected $p = 0.031$, Holm-corrected $p = 0.406$). Descriptive accuracy differences between language conditions ranged from -5.0 to $+4.5$ percentage points across models.

Table 3. McNemar's exact tests comparing English and Turkish modal responses within each model.

Model	Discordant (TRK ✓/ENG ✗)	Discordant (ENG ✓/TRK ✗)	p (Exact)	Holm–Bonferroni
Gemini 3	6	0	0.031	0.406
ChatGPT-5.2	2	2	1.000	1.000
Grok 4.1	5	3	0.727	1.000
Gemini 3 Pro	1	1	1.000	1.000
ChatGPT-5.2 Pro	2	3	1.000	1.000
SuperGrok 4.1	3	3	1.000	1.000

Computed on items with a clear modal response (≥ 4 of 6) in both language versions, yielding 98, 91, 92, 94, 96, and 95 retained items for Gemini 3, ChatGPT-5.2, Grok 4.1, Gemini 3 Pro, ChatGPT-5.2 Pro, and SuperGrok 4.1, respectively. Holm–Bonferroni correction was applied across the full family of 14 tests. ✓, item answered correctly; ✗, item answered incorrectly.

3.4. Item-Level Cross-Language Equivalence

At the level of individual items, mean item difficulty was virtually identical between language versions (0.806 in Turkish versus 0.807 in English; mean difference -0.0003 across 100 items). This full-response comparison, which uses every individual response rather than modal answers, showed no significant cross-language difference (Wilcoxon signed-rank $p = 0.348$; item-clustered bootstrap 95% CI for the mean difference -2.3 to $+2.4$ percentage points). Item-difficulty parameters were strongly correlated between versions (Pearson $r = 0.928$, 95% CI 0.870–0.966; Spearman $\rho = 0.791$), indicating that both the rank order and the magnitude of item difficulty were preserved across languages. Per-model cross-language correlations of item difficulty ranged from $r = 0.696$ to $r = 0.885$, and per-model Cohen's κ between Turkish and English modal responses ranged from 0.891 to 0.973, with 91% to 98% of items receiving identical modal responses across language versions (Table 4).

Table 4. Item-level cross-language equivalence per model: correlation of item difficulty and agreement of modal responses between the Turkish and English versions.

Model	Tier	Pearson r	Spearman ρ	κ (Modal)	Modal Concordance (%)
Gemini 3	Free	0.803	0.797	0.910	93 (91/98)
ChatGPT-5.2	Free	0.880	0.790	0.931	95 (86/91)
Grok 4.1	Free	0.696	0.660	0.891	91 (84/92)
Gemini 3 Pro	Paid	0.847	0.694	0.973	98 (92/94)
ChatGPT-5.2 Pro	Paid	0.885	0.857	0.935	95 (91/96)
SuperGrok 4.1	Paid	0.795	0.708	0.921	94 (89/95)

Item-difficulty correlations (Pearson r , Spearman ρ) were computed on the per-item proportion correct across six runs per language for all 100 items. The κ (modal) and modal-concordance columns were computed only on items with a clear modal response (≥ 4 of 6) in both language versions, yielding 91 to 98 retained items per model. Pooled across all 100 items, mean item difficulty was 0.806 (Turkish) versus 0.807 (English); pooled Pearson $r = 0.928$ (95% CI 0.870–0.966), Spearman $\rho = 0.791$. Modal concordance shown as % (concordant/retained items). κ , Cohen’s kappa.

3.5. Subscription-Tier Differences

McNemar’s exact tests comparing free and paid variants within each developer dyad did not reach statistical significance after Holm–Bonferroni correction under either language condition (Table 5). At the aggregate run level, a nominal English-condition advantage for paid variants did not survive correction, and no significant difference emerged in the Turkish condition. Although paid variants tended to show marginally higher accuracy within developer-matched pairs, the differences were incremental rather than indicative of categorically different levels of specialist knowledge.

Table 5. McNemar’s exact tests comparing free and paid model variants within each developer dyad, by language version.

Developer Dyad	Language	Discordant (Paid ✓/Free ✗)	Discordant (Free ✓/Paid ✗)	p (Exact)	Holm– Bonferroni
Gemini 3/Gemini 3 Pro	English	7	2	0.180	1.000
Gemini 3/Gemini 3 Pro	Turkish	3	0	0.250	1.000
ChatGPT-5.2/ChatGPT-5.2 Pro	English	9	2	0.065	0.785
ChatGPT-5.2/ChatGPT-5.2 Pro	Turkish	6	1	0.125	1.000
Grok 4.1/SuperGrok 4.1	English	5	3	0.727	1.000
Grok 4.1/SuperGrok 4.1	Turkish	2	3	1.000	1.000

Discordant pairs denote items answered correctly by one variant but not the other (b, paid correct/free incorrect; c, free correct/paid incorrect). Computed on items with a clear modal response (≥ 4 of 6) in both variants, yielding 97, 94, and 97 retained items in English and 96, 92, and 90 in Turkish for the Gemini, ChatGPT, and Grok dyads. Holm–Bonferroni correction was applied across the full family of 14 tests. At the aggregate run level, Welch’s t -test comparing free and paid runs gave $p = 0.004$ (English; Holm-corrected $p = 0.062$) and $p = 0.088$ (Turkish; Holm-corrected $p = 0.971$). ✓, item answered correctly; ✗, item answered incorrectly.

3.6. Structure of Residual Divergences

Across all 72 model runs, 32 of the 100 items were answered correctly by every model, in both languages, in every one of the six runs; applying the modal-response criterion of Section 2.5 instead, the corresponding figure is 59 of 100. Most remaining items were close to ceiling. Divergences were not randomly distributed: on the most difficult items, the same single response tended to dominate across models in both the English and Turkish versions. The few divergent items were concentrated on highly specialized, reference-dependent, and doctrine-sensitive content. Such items included questions turning on precise numeric values, such as the ratio of type I to type III collagen in adults; on “most common” epidemiological claims, such as the leading cause of death in paraplegic patients with chronic pressure ulcers; and on questions whose correct option depends on the timing

considered, such as the most common complication after otoplasty, which differs according to whether early (hematoma) or late complications (recurrence) are taken into account.

4. Discussion

This study evaluated six contemporary LLMs on the national board examination of the TSPRAS, administered in parallel Turkish and English versions across six independent runs per condition. Four findings emerged: all models performed above the specialist passing threshold in the large majority of administrations, responses were highly reproducible across repeated administrations, performance did not differ significantly between the Turkish and English examination versions, and paid variants offered only marginal advantages over their free counterparts.

The passing-level performance observed here extends the strong results previously reported for LLMs on general medical examinations [6–10] to a national plastic surgery board examination and is consistent with findings from specialty-specific plastic surgery assessments, including the PSITE [11–21]. That contemporary models answer specialist questions at the passing level is, by itself, no longer surprising. The more relevant observation for surgical education is that this performance was achieved on examination content that is not publicly available, which speaks directly to a limitation of many prior studies. Because the items are not in the public domain, direct exposure through verbatim memorization of these specific items is unlikely, so this form of contamination is unlikely to explain the observed accuracy. Complete exclusion of contamination is nonetheless unattainable for deployed commercial models, as it would require items authored *de novo* on a system with no external network connection and administered immediately so that they could not have entered a training corpus beforehand. Indirect exposure cannot be ruled out either, as the underlying curriculum content is covered by standard textbooks and by publicly circulating items from other examinations.

By drawing on a national board examination outside the United States and in a language other than English, this work broadens an evidence base that has been narrow and largely anglophone. It provides, to our knowledge, the first evaluation of LLMs on a national plastic surgery board examination delivered in two official parallel language versions, addressing the underrepresentation of other countries, curricula, and languages.

An important finding from an educational perspective is the absence of a strong language effect. Performance did not differ significantly between the Turkish and English examination versions after correction, and item difficulty was strongly correlated across languages, indicating that the same items were comparably difficult irrespective of examination language. This contrasts with the well-documented reduction in human examination performance when testing occurs in a second rather than a first language [22–25] and converges with a small but growing body of cross-language LLM evaluations. The L1 and L2 distinction is not mechanistically transferable to these systems. Although English is likely overrepresented both in the pretraining corpora and in the subsequent alignment of current models, this dominance is not equivalent to a human native language: these systems do not acquire language sequentially and have no mother tongue but are optimized over multilingual data of unequal size and quality. Any residual language effect is therefore more plausibly attributable to differences in training-data composition than to an order of acquisition analogous to human L1 and L2.

Yao et al. evaluated seven LLMs trained on English and Chinese corpora using the Chinese National Medical Licensing Examination (CNMLE) and found that English-trained models performed comparably to Chinese-trained models, with only a slight advantage for the latter, suggesting that performance differences are driven less by language than by training-data composition [30].

Convergent evidence comes from Wójcik et al., who evaluated three LLMs on parallel English and Polish versions of a national medical examination across multiple runs per language. Two of the three models showed no significant cross-language difference, while one showed a significant difference between versions. That this pattern holds across Polish, a West Slavic language, and Turkish, a Turkic language, suggests further investigation across additional language families is warranted [31].

Recent work has further disentangled question language from prompt language. Strasser et al. evaluated five LLMs on identical medical multiple-choice questions in German, French, and Italian, while varying whether instructions were given in English or matched the question language. Across all models, prompt language influenced accuracy more than question language itself, with English prompts generally yielding higher performance. Because prompts in the present study were held constant in English across both question conditions, the absence of a significant question-language effect is consistent with the modest question-language effects reported under comparable English-prompt conditions [32].

For surgical education and education in general, the limited influence of question language observed in this study is practically encouraging, as it suggests that trainees preparing in a language other than English are not necessarily disadvantaged when using these tools. However, this observation applies specifically to the language of the examination content, as prompt language was held constant, and it may exert a stronger effect on performance.

Subscription tier had little effect on measured performance. Within each developer-matched pair, paid variants showed only marginal and non-significant advantages over their free counterparts, suggesting that for the retrieval of factual specialist knowledge assessed by a board-style examination, the paid tier adds little. A plausible interpretation is that the residual distance from perfect accuracy reflects genuinely demanding content rather than a tier-dependent capability gap: the few items on which models consistently diverged concentrated on highly specialized, reference-dependent, and doctrine-sensitive content, and this pattern persisted across both free and paid variants. The advantages of premium tiers may instead emerge for more complex tasks, such as extended multi-step reasoning or code generation, rather than for factual knowledge retrieval; evaluating that possibility lies beyond the scope of the present study. Several explanations are consistent with this pattern. The single-letter response format cannot capture any paid-tier advantage in reasoning depth or the handling of ambiguity, and near-ceiling accuracy leaves little room for tier differences on the hardest items. Moreover, for contemporary consumer products, a paid tier is largely defined by usage-related and feature-related benefits (for example, higher usage limits, additional modalities, and earlier access to new models) rather than by a categorically stronger factual-knowledge base, and single-best-answer (SBA) factual recall is the simplest capability on which free and paid tiers have most nearly converged. This negative finding therefore applies to board-style factual retrieval as measured here rather than to free and paid tiers in general. From an educational-equity perspective, the comparable performance of free and paid variants is nonetheless reassuring, as it suggests that trainees and departments with limited resources are not disadvantaged for this particular use.

These findings should be interpreted with the nature of the assessment in mind. An SBA examination probes declarative and applied knowledge rather than clinical judgment, operative skill, or the visual and aesthetic reasoning central to plastic surgery, so strong performance should not be read as evidence of clinical competence. The combination of high accuracy and high inter-run reliability on core content with persistent divergence on a small set of under-determined items is itself a source of risk: stable, confidently phrased

outputs may convey a reliability the underlying knowledge does not always justify, and the principal hazard for a trainee is not the obviously wrong answer but the fluent, plausible, yet incomplete or contextually inappropriate one. These systems are therefore best used as supportive but fallible study aids within a human-in-the-loop (HITL) model, with expert oversight. Because clinicians increasingly consult these tools at the point of care, reliable command of specialist knowledge is a necessary but not sufficient condition for safe clinical use. This study assessed answer selection on SBA items only and did not assess explanation quality, calibration, hallucination rate, teaching value, or clinical judgment, so it supports no claim about clinical decision support or point-of-care consultation. Within these limits, the findings support cautious, supervised use of LLMs as study aids rather than unsupervised clinical deployment.

Several limitations should be acknowledged. The examination is text-based and therefore captures declarative and applied knowledge but not the visual, spatial, and procedural reasoning intrinsic to plastic surgery, so the present results should not be extrapolated to image-based or operative assessment. The evaluation was limited to a single examination cycle and to the two examination languages tested, with prompts held constant in English, and was conducted through the official public web interfaces, which preclude control over decoding parameters and version locking and do not expose the underlying model version; the tier contrast therefore compares user-facing products rather than architectures, and item scoring followed the official answer key. The findings should therefore be interpreted as a snapshot assessment of a selection of widely used models representative of the contemporary LLM landscape as available during the study period. Finally, the analysis is based on observable response behavior and does not characterize the reasoning processes underlying correct or incorrect responses.

5. Conclusions

In conclusion, contemporary LLMs answered specialized board examination questions at a passing level in a demanding medical domain, here the national plastic surgery board examination of the TSPRAS, with high reproducibility, no significant difference between the Turkish and English examination versions, and only marginal differences between free and paid variants. For specialist medical education, this supports their use as study aids on the tested multiple-choice format under a constant English prompt, including for trainees preparing in languages other than English and in resource-limited settings, provided they are used as supportive but fallible tools under specialist supervision rather than equated with clinical competence. The few divergent items, concentrated on highly specialized, non-canonical content and not closed by paid subscriptions, mark the current limits of these tools. Future work should move beyond aggregated board scores to clinically relevant, subspecialty-specific tasks, including multimodal image interpretation, operative planning, and decision support, while recognizing that ground truth in specialist medicine is not always universal and may vary across curricula, textbooks, and established schools of thought.

Author Contributions: Conceptualization, I.G. and C.D.; methodology, I.G., H.S. and C.D.; software, I.G. and H.S.; validation, I.G., H.S. and C.D.; formal analysis, I.G. and H.S.; investigation, I.G. and H.S.; resources, C.D.; data curation, I.G., U.C. and H.S.; writing—original draft preparation, I.G. and H.S.; writing—review and editing, I.G., H.S., C.D., A.K., U.C. and G.G.; visualization, I.G.; supervision, C.D.; project administration, I.G., H.S. and C.D.; funding acquisition, none. 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.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data are available from the corresponding author upon request. The examination questions cannot be made publicly available due to institutional restrictions.

Acknowledgments: The authors thank the Turkish Society of Plastic, Reconstructive and Aesthetic Surgeons (TSPRAS) for providing the examination materials. Generative AI-based tools (Claude Opus 4.6) were used during manuscript preparation to support the authors in auxiliary tasks, including language editing and grammar and stylistic refinement, as well as assistance with code generation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
ANI	artificial narrow intelligence
API	application programming interface
ASPS	American Society of Plastic Surgeons
CI	confidence interval
CNMLE	Chinese National Medical Licensing Examination
GenAI	generative artificial intelligence
GPAI	general-purpose AI
HITL	human-in-the-loop
L1	first language
L2	second language
LLM	large language model
MCQ	multiple-choice question
MLLM	multimodal large language model
OSCE	Objective Structured Clinical Examination
PSITE	Plastic Surgery In-Training Examination
RAG	retrieval-augmented generation
SBA	single-best-answer
SD	standard deviation
TSPRAS	Turkish Society of Plastic, Reconstructive and Aesthetic Surgeons
κ	Cohen's kappa

References

1. Kiwan, O.; Al-Kalbani, M.; Rafie, A.; Hijazi, Y. Artificial intelligence in plastic surgery, where do we stand? *JPRAS Open* **2024**, *42*, 234–243. [[CrossRef](#)] [[PubMed](#)] [[PubMed Central](#)]
2. Lewis, C.J.; Reid, A.J. The evolving integration of artificial intelligence in plastic surgery. *J. Plast. Reconstr. Aesthetic Surg.* **2025**, *110*, 361–362. [[CrossRef](#)] [[PubMed](#)]
3. Bhargava, V.; Kar, A.; Gupta, S.; Yadav, C.; Singh, K.; Lata, P. AI in layman's life. *AI Ethics* **2026**, *6*, 97. [[CrossRef](#)]
4. Szadeczky, T.; Bederna, Z. Risk, regulation, and governance: Evaluating artificial intelligence across diverse application scenarios. *Secur. J.* **2025**, *38*, 35. [[CrossRef](#)]
5. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [[CrossRef](#)] [[PubMed](#)]
6. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [[CrossRef](#)] [[PubMed](#)] [[PubMed Central](#)]

7. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. Erratum in *JMIR Med. Educ.* **2024**, *10*, e57594. <https://doi.org/10.2196/57594>. [CrossRef] [PubMed] [PubMed Central]
8. Brin, D.; Sorin, V.; Vaid, A.; Soroush, A.; Glicksberg, B.S.; Charney, A.W.; Nadkarni, G.; Klang, E. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci. Rep.* **2023**, *13*, 16492. [CrossRef] [PubMed] [PubMed Central]
9. Katz, U.; Cohen, E.; Shachar, E.; Somer, J.; Fink, A.; Morse, E.; Shreiber, B.; Wolf, I. GPT versus resident physicians: A benchmark based on official board scores. *NEJM AI* **2024**, *1*, AIdbp2300192. [CrossRef]
10. Newton, P.; Xiromeriti, M. ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. *Assess. Eval. High. Educ.* **2024**, *49*, 781–798. [CrossRef]
11. Najafali, D.; Reiche, E.; Araya, S.; Orellana, M.; Liu, F.C.; Camacho, J.M.M.; Patel, S.A.; Broyles, J.M.; Dorafshar, A.H.; Morrison, S.D.; et al. Artificial Intelligence Augmentation: Performance of GPT-4 and GPT-3.5 on the Plastic Surgery In-service Examination. *Plast. Reconstr. Surg. Glob. Open* **2025**, *13*, e6645. [CrossRef] [PubMed] [PubMed Central]
12. Hubany, S.S.; Scala, F.D.; Hashemi, K.; Kapoor, S.B.; Fedorova, J.R.B.; Vaccaro, M.J.B.; Ridout, R.P.B.; Hedman, C.C.B.; Kellogg, B.C.; Barone, A.A.L. ChatGPT-4 Surpasses Residents: A Study of Artificial Intelligence Competency in Plastic Surgery In-service Examinations and Its Advancements from ChatGPT-3.5. *Plast. Reconstr. Surg. Glob. Open* **2024**, *12*, e6136. [CrossRef] [PubMed] [PubMed Central]
13. Najafali, D.; Reiche, E.; Araya, S.; Camacho, J.M.; Liu, F.C.; Johnstone, T.; A Patel, S.; Morrison, S.D.; Dorafshar, A.H.; Fox, P.M. Bard Versus the 2022 American Society of Plastic Surgeons In-Service Examination: Performance on the Examination in Its Intern Year. *Aesthetic Surg. J. Open Forum* **2023**, *6*, ojad066. [CrossRef] [PubMed] [PubMed Central]
14. Raine, B.E.; Kozlowski, K.A.; Fowler, C.C.; Frey, J.D. Performance of ChatGPT on the Plastic Surgery In-Training Examination. *Eplasty* **2024**, *24*, e68. [PubMed] [PubMed Central]
15. Gupta, R.; Herzog, I.; Park, J.B.; Weisberger, J.; Firouzbakht, P.; Ocon, V.; Chao, J.; Lee, E.S.; A Mailey, B. Performance of ChatGPT on the Plastic Surgery Inservice Training Examination. *Aesthet. Surg. J.* **2023**, *43*, NP1078–NP1082. Erratum in *Aesthet. Surg. J.* **2024**, *44*, 232. <https://doi.org/10.1093/asj/sjad316>. [CrossRef] [PubMed]
16. Humar, P.; Asaad, M.; Bengur, F.B.; Nguyen, V. ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination. *Aesthet. Surg. J.* **2023**, *43*, NP1085–NP1089. Erratum in *Aesthet. Surg. J.* **2024**, *44*, 232. <https://doi.org/10.1093/asj/sjad316>. [CrossRef] [PubMed]
17. Copeland-Halperin, L.R.; O'Brien, L.; Copeland, M. Evaluation of Artificial Intelligence-generated Responses to Common Plastic Surgery Questions. *Plast. Reconstr. Surg. Glob. Open* **2023**, *11*, e5226. [CrossRef] [PubMed] [PubMed Central]
18. Gupta, R.; Park, J.B.; Herzog, I.; Yosufi, N.; Mangan, A.; Firouzbakht, P.K.; Mailey, B.A. Applying GPT-4 to the Plastic Surgery Inservice Training Examination. *J. Plast. Reconstr. Aesthet. Surg.* **2023**, *87*, 78–82. [CrossRef] [PubMed]
19. DiDonna, N.; Shetty, P.N.; Khan, K.; Damitz, L. Unveiling the Potential of AI in Plastic Surgery Education: A Comparative Study of Leading AI Platforms' Performance on In-training Examinations. *Plast. Reconstr. Surg. Glob. Open* **2024**, *12*, e5929. [CrossRef] [PubMed] [PubMed Central]
20. Hsieh, C.H.; Hsieh, H.Y.; Lin, H.P. Evaluating the performance of ChatGPT-3.5 and ChatGPT-4 on the Taiwan plastic surgery board examination. *Heliyon* **2024**, *10*, e34851. [CrossRef] [PubMed] [PubMed Central]
21. Lei, Y.H.; Chen, C.C.; Shen, C.J. Token-splitting improves GPT-4.1 performance on plastic surgery exams: Implications for AI-Assisted medical education. *Med. Educ. Online* **2025**, *30*, 2602788. [CrossRef] [PubMed] [PubMed Central]
22. Bernhofer, J.; Tonin, M. The effect of the language of instruction on academic performance. *Labour Econ.* **2022**, *78*, 102218. [CrossRef]
23. Vander Beken, H.; Brysbaert, M. Studying texts in a second language: The importance of test type. *Biling. Lang. Cogn.* **2018**, *21*, 1062–1074. [CrossRef]
24. Desselle, S.P.; Shane, P.A. Native English speakers and English as a Foreign Language (EFL) students' performance and notetaking in a Doctor of Pharmacy health systems course. *Res. Soc. Adm. Pharm.* **2019**, *15*, 1154–1159. [CrossRef] [PubMed]
25. Haag, N.; Heppt, B.; Stanat, P.; Kuhl, P.; Pant, H.A. Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learn. Instr.* **2013**, *28*, 24–34. [CrossRef]
26. Turkish Society of Plastic Reconstructive and Aesthetic Surgery (TSPRAS). Yeterlik Sınavı (Board Certification Examination Information). Available online: <https://dernek.plastikcerrahi.org.tr/menu/24/yeterlik-sinavi> (accessed on 8 June 2026).
27. Turkish Society of Plastic Reconstructive and Aesthetic Surgery (TSPRAS). Yeterlik Sınavı Alt Yönergesi (Board Certification Examination Regulations). Available online: <https://dernek.plastikcerrahi.org.tr/menu/29/yeterlik-sinavi-alt-yonergesi> (accessed on 8 July 2026).
28. Siam, M.K.; Varela, A.; Faruk, M.J.H.; Cheng, J.Q.; Gu, H.; Al Maruf, A.; Aung, Z. Benchmarking large language models on the United States medical licensing examination for clinical reasoning and medical licensing scenarios. *Sci. Rep.* **2025**, *16*, 1387. [CrossRef] [PubMed] [PubMed Central]

29. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
30. Yao, Z.; Duan, L.; Xu, S.; Chi, L.; Sheng, D. Performance of Large Language Models in the Non-English Context: Qualitative Study of Models Trained on Different Languages in Chinese Medical Examinations. *JMIR Med. Inf.* **2025**, *13*, e69485. [[CrossRef](#)] [[PubMed](#)] [[PubMed Central](#)]
31. Wójcik, D.; Adamiak, O.; Czerepak, G.; Tokarczuk, O.; Szalewski, L. A bi-linguistic comparative analysis of ChatGPT-4, Gemini, and Claude performance on Polish medical-dental final examinations. *Sci. Rep.* **2025**, *15*, 33083. Erratum in *Sci. Rep.* **2025**, *15*, 43849. <https://doi.org/10.1038/s41598-025-31400-8>. [[CrossRef](#)] [[PubMed](#)] [[PubMed Central](#)]
32. Strasser, L.M.; Anschuetz, W.; Dennstädt, F.; Hastings, J. Performance Evaluation of Large Language Models in Multilingual Medical Multiple-Choice Questions: Mixed Methods Study. *JMIR Med. Educ.* **2026**, *12*, e81399. [[CrossRef](#)] [[PubMed](#)] [[PubMed Central](#)]

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.