

Accuracy of GPT-3.5, GPT-4.0, and GPT-5.0 in Neurosurgery-Related Questions: A Comparative Study Including Self-Correction Analysis

Leo Heric Caxias Gomes

Federal University of Amazonas (UFAM)

Frida Chaves Giordani

University of the State of Amazonas

Giovana dos Santos Couto

Federal University of Amazonas (UFAM)

Isabella Christine Lima Maia Lira

Federal University of Amazonas (UFAM)

Paulo Guilherme Mattos Edwards

Federal University of Amazonas (UFAM)

Robson Luis Oliveira Amorim

amorim.robson@gmail.com

University of São Paulo (USP)

Method Article

Keywords: artificial intelligence, large language models, neurosurgery, GPT-3.5, GPT-4.0, GPT-5.0, accuracy, self-correction

Posted Date: June 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9751851/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Introduction:

Artificial intelligence (AI), and large language models (LLMs) in particular, are increasingly integrated into clinical and educational settings across medicine. This study compares the accuracy of three successive GPT model versions—GPT-3.5 (Standard), GPT-4.0 (Standard), and GPT-5.0 (Standard)—on neurosurgery-related multiple-choice questions administered in both Portuguese and English.

Objective

To compare initial response accuracy and self-correction capacity across GPT-3.5, GPT-4.0, and GPT-5.0, using a standardized neurosurgical question bank in two languages.

Methods

This descriptive-comparative study used a 250-item multiple-choice question bank curated by two independent neurosurgeons, organized into five difficulty levels (50 items per level). Each model answered every question once in Portuguese and once in English (1,500 total evaluations). Accuracy was determined against a pre-established consensus answer key. Following each incorrect initial response, a standardized revision prompt was issued (“This is not the correct answer, try again”) and the final response was recorded. Grouped logistic regression with item-level clustering and paired McNemar tests were used for inferential analyses.

Results

GPT-5.0 demonstrated the highest overall accuracy (78.0% combined across languages), compared with GPT-3.5 (61.8%) and GPT-4.0 (60.8%). In clustered logistic regression, GPT-5.0 showed significantly higher odds of an initial correct response compared with GPT-3.5 ($OR \approx 2.24$, $p < 0.001$), while GPT-4.0 did not differ significantly from GPT-3.5 ($p = 0.674$). No significant difference in accuracy was observed between Portuguese and English prompts for any model (McNemar $p = 0.311$, 0.146 , and 0.873 for GPT-3.5, GPT-4.0, and GPT-5.0, respectively). Among initially incorrect responses, GPT-5.0 showed significantly higher self-correction rates than GPT-3.5 ($OR \approx 5.4$, $p < 0.001$), while GPT-4.0 showed a modest but significant improvement ($OR \approx 2.2$, $p = 0.046$).

Conclusion

GPT-5.0 represents a meaningful advance over its predecessors in neurosurgery-related knowledge, with language-independent performance and substantially superior self-correction capacity. However, overall

accuracy remains below 80%, and the absence of human benchmarking and the single-trial design limit direct clinical applicability. These findings underscore both the potential and the current limitations of LLMs in specialized medical education.

INTRODUCTION

Artificial intelligence (AI) encompasses computational systems designed to replicate cognitive functions such as learning, reasoning, and problem-solving. Among its most transformative recent manifestations are large language models (LLMs), which process and generate natural language at scale [1, 2]. These systems have been rapidly integrated across multiple domains of medicine, where they show promise in supporting diagnosis, clinical documentation, and medical education [3, 4].

LLMs are frequently positioned as potential disruptors in healthcare, given their ability to process large volumes of unstructured data and identify patterns beyond human capacity [5]. However, these capabilities are accompanied by well-documented limitations, including the generation of plausible but factually incorrect statements—a phenomenon known as “hallucination”—and performance degradation on tasks requiring complex multi-step reasoning or specialized expertise [6, 7].

Studies evaluating GPT-3.5 on medical evidence summarization have found that automated performance metrics frequently fail to capture output quality, and that models can produce inconsistent or factually unsupported statements [8]. Comparative analyses of GPT-4.0 demonstrated near-human-level performance across several general domains, yet clinical applications revealed persistent vulnerabilities. For instance, when tested using the United Kingdom and Ireland Training Examination as a surrogate for the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedics) Section 1 examination, ChatGPT achieved only 35.8%—approximately 30 percentage points below the passing grade—with markedly low performance on trauma questions [9]. In the specific field of neurosurgery, Daungsupawong and Wiwanitkit [10] found that the AI model ranked ninth of eleven participants when answering questions generated by neurosurgery residents, underscoring the gap between general and domain-specific performance.

GPT-5.0 introduces architectural improvements reported to include expanded training data, enhanced chain-of-thought reasoning, and reduced hallucination rates [11, 12]. Preliminary evaluations on health-related benchmarks suggest performance gains over predecessor models, though clinical safety concerns remain [13]. The extent to which these improvements translate to neurosurgical content—a domain requiring integration of anatomical, physiological, and clinical knowledge—has not been systematically investigated.

A further unresolved question concerns language equity. Current LLM training datasets are predominantly English-language, potentially creating performance disparities when models are queried in other languages [7]. For non-English-speaking healthcare environments, such as Brazil, this issue has direct implications for clinical and educational utility.

The present study addresses these gaps by systematically comparing the accuracy of GPT-3.5, GPT-4.0, and GPT-5.0 on a standardized neurosurgical question bank administered in both Portuguese and English. A secondary aim is to evaluate the capacity of each model to self-correct an initially incorrect response when prompted.

METHODS

Study Design

This is a descriptive-comparative study evaluating the accuracy of three GPT model versions on a structured neurosurgical question bank. The comparative design allowed systematic analysis of performance differences across model versions, prompt languages, and item difficulty levels.

Question Bank Development and Reference Standard

A 250-item multiple-choice question bank was constructed from established neurosurgery textbooks, peer-reviewed educational references, and widely used specialty examination preparatory materials. All questions followed a single-best-answer format with four alternatives. Items were independently reviewed by two neurosurgeons, who assessed clinical relevance, wording clarity, and alignment with accepted current knowledge. Disagreements were resolved through structured discussion until a consensus answer was reached. The resulting consensus answer key served as the reference standard (gold standard) for all accuracy assessments.

Items were classified into five difficulty levels (Level 1 through Level 5) a priori, based on the conceptual domain and cognitive complexity required: Level 1 items addressed foundational anatomical and physiological principles; Levels 2 and 3 addressed clinical recognition and standard management; Levels 4 and 5 addressed advanced clinical decision-making, rare conditions, and nuanced evidence appraisal. This classification was agreed upon by both reviewing neurosurgeons prior to model testing. Although formal psychometric validation of item difficulty was not performed—a limitation acknowledged below—the classification scheme was determined prior to model evaluation and applied uniformly across all items.

Question Set Details

The question bank comprised 250 items, with 50 allocated to each of the five difficulty levels. All items were presented as multiple-choice questions with four alternatives and a single correct answer. Each item was administered in both Portuguese and English, yielding 500 item-language combinations per model and 1,500 total observations across the three models.

Model Configuration and Evaluation Conditions

Three GPT model versions were evaluated: GPT-3.5 (Standard), GPT-4.0 (Standard) and GPT-5.0 (Standard). All models were accessed via their respective default configurations through the OpenAI platform at the time of evaluation. No modifications to temperature, system-level context, custom instructions, memory augmentation, retrieval-augmented generation, or additional prompt engineering were applied. The evaluation was designed to reflect real-world “out-of-the-box” conditions. Each question was submitted once per model per language; this single-trial design precludes estimation of intra-model variability and is acknowledged as a methodological limitation.

Revision Request Protocol

Following each initial model response, the response was classified as correct or incorrect by comparison with the reference answer key. When the initial response was incorrect, a standardized revision prompt was issued in the same language as the original question: “This is not the correct answer, try again.” The model’s subsequent response was then recorded as the final response. This protocol was applied uniformly across all models and languages. Because the revision prompt does not specify which element of the response was incorrect and does not provide additional information, it functions as a generic recall cue, enabling assessment of each model’s capacity for autonomous error correction.

For initially correct responses, no revision prompt was issued, and the initial response was carried forward as the final response. Two binary indicators were derived from this protocol: (1) whether the answer changed after the revision prompt (changed: yes/no), and (2) whether the final response was correct (correct_final: yes/no). Analyses of self-correction effectiveness were restricted to the subset of initially incorrect responses.

Accuracy Assessment

Each model response was compared with the pre-established reference answer key by two evaluators. Responses were classified as correct (1) or incorrect (0) based on exact correspondence with the consensus answer. In cases of ambiguous or free-text responses, a structured adjudication protocol was applied: the response was first assessed for direct selection of the correct alternative; if the model elaborated on the correct answer without explicitly naming it, the response was reviewed by both evaluators and classified by consensus. Discrepancies in scoring were resolved by discussion.

Statistical Analysis

Descriptive statistics were used to summarize initial accuracy, final accuracy, and correction rates across models, difficulty levels, and languages. Inferential analyses were conducted to test predefined hypotheses regarding model performance, language effects, and revision effects.

The unit of analysis was the individual item-language-model observation. Each item was assigned a composite identifier combining difficulty level and item number (1–50). Each item was evaluated under six conditions (three models × two languages). Logistic regression models with robust standard errors

clustered at the item level were fitted to account for within-item correlation arising from repeated evaluation of the same question across models and languages. An initial model included main effects for model and difficulty level; a subsequent model incorporated model-by-difficulty interaction terms to assess whether accuracy differences between models varied by difficulty. Results were expressed as odds ratios (OR) with 95% confidence intervals. Adjusted predicted probabilities were derived from marginal effects.

Language effects were assessed using McNemar’s test for paired comparisons between Portuguese and English prompts, stratified by model, with exact p-values reported. Among initially incorrect responses, the effect of the revision prompt on accuracy was modeled using logistic regression with item-level clustering, with model, difficulty level, and prompt language as covariates. A reduced model including model as the sole predictor was used to generate a forest plot of odds ratios. Mean accuracy gain (final minus initial accuracy) was computed and visualized as heatmaps stratified by model, difficulty level, and language. All tests were two-sided; $p < 0.05$ was considered statistically significant. All analyses were conducted using Stata version 16.0 (StataCorp, College Station, TX).

Ethical Considerations

This study involved no human participants, biological specimens, or identifiable personal data. All data were generated by querying publicly accessible AI platforms using an investigator-developed question bank. In accordance with national and institutional guidelines, formal ethics committee review was not required for studies of this nature. The study was conducted in accordance with the principles of academic integrity and transparency in AI research.

RESULTS

Overview

A total of 1,500 item-level observations were analyzed, corresponding to 250 unique items evaluated across three language models (GPT-3.5, GPT-4.0, and GPT-5.0) and two prompt languages (Portuguese and English). Of 1,500 initial responses, 498 were incorrect and thus eligible for the revision analysis.

Initial Accuracy: Overall and by Language

Table 1 presents initial accuracy stratified by model, language, and difficulty level. Overall accuracy across both languages was 61.8% for GPT-3.5, 60.8% for GPT-4.0, and 78.0% for GPT-5.0. For GPT-3.5, accuracy was slightly higher in Portuguese (63.6%) than in English (60.0%). For GPT-4.0, the direction was reversed, with higher accuracy in English (63.2%) than in Portuguese (58.4%). GPT-5.0 showed identical accuracy in both languages (78.0% for both).

Table 1
Initial accuracy (%) by model, language, and difficulty level

Model / Language	Level 1	Level 2	Level 3	Level 4	Level 5	Overall (n = 250)
GPT-3.5 / Portuguese	38%	80%	64%	66%	70%	63.6%
GPT-3.5 / English	36%	78%	80%	54%	52%	60.0%
GPT-4.0 / Portuguese	38%	68%	66%	64%	56%	58.4%
GPT-4.0 / English	38%	74%	76%	66%	62%	63.2%
GPT-5.0 / Portuguese	82%	86%	84%	76%	62%	78.0%
GPT-5.0 / English	76%	80%	86%	74%	74%	78.0%

Accuracy (%) represents the proportion of initial correct responses per 50 items. Overall accuracy is pooled across both languages.

Accuracy by Difficulty Level

A notable and consistent finding across GPT-3.5 and GPT-4.0 was that Level 1 items yielded the lowest accuracy of all levels, with both models achieving only 37–38% at Level 1 combined across languages—lower than their performance at Levels 2 through 5. For GPT-5.0, this pattern was substantially attenuated, with Level 1 accuracy reaching 79% (combined), consistent with the generally higher baseline of that model. For GPT-3.5 and GPT-4.0, accuracy peaked at Levels 2 and 3 before declining at Levels 4 and 5, suggesting that foundational questions classified as Level 1 posed particular challenges for earlier model versions. GPT-5.0 showed a monotonic decline from Level 3 (85% combined) to Level 5 (68% combined), consistent with the expected gradient of difficulty for that model.

In clustered logistic regression adjusted for difficulty level, model performance differed significantly across versions (Wald $\chi^2 = 53.08$, $p < 0.001$). Compared with GPT-3.5, GPT-5.0 demonstrated significantly higher odds of providing a correct initial response (OR ≈ 2.24 , 95% CI [to be reported]; $p < 0.001$), whereas GPT-4.0 did not differ significantly from GPT-3.5 ($p = 0.674$). When model-by-difficulty interaction terms were included, the interaction was statistically significant (Wald $\chi^2 = 71.95$, $p < 0.001$), indicating that model differences varied by difficulty level. GPT-5.0 showed the largest advantage over GPT-3.5 at Level 1, with this advantage progressively attenuating at higher levels. GPT-4.0 did not differ significantly from GPT-3.5 at any individual difficulty level. Adjusted predicted probabilities are presented in Fig. 1.

Language Effects

Paired comparisons between Portuguese and English prompts using McNemar’s test revealed no statistically significant differences in accuracy for any model. McNemar p-values were 0.311 for GPT-3.5, 0.146 for GPT-4.0, and 0.873 for GPT-5.0. Differences in proportions between languages were small in magnitude and inconsistent in direction (GPT-3.5: +3.6 percentage points favoring Portuguese; GPT-4.0:

–4.8 pp favoring English; GPT-5.0: 0.0 pp). These findings indicate that prompt language did not significantly influence accuracy in any of the three models.

Self-Correction After Revision Prompt

Among the 498 initially incorrect responses, a revision prompt (“This is not the correct answer, try again”) was issued following each error. Table 2 summarizes the self-correction outcomes.

Table 2
Self-correction outcomes among initially incorrect responses, by model and language

Model	Language	Wrong (n)	Changed answer (n / %)	Corrected (n)	Correction rate ¹ (%)	Correction rate ² (%)
GPT-3.5	Portuguese	91	17 (18.7%)	5	29.4%	5.5%
GPT-3.5	English	100	13 (13.0%)	4	30.8%	4.0%
GPT-4.0	Portuguese	104	25 (24.0%)	12	48.0%	11.5%
GPT-4.0	English	92	22 (23.9%)	8	36.4%	8.7%
GPT-5.0	Portuguese	55	22 (40.0%)	13	59.1%	23.6%
GPT-5.0	English	56	17 (30.4%)	8	47.1%	14.3%

¹ Correction rate among those who changed their answer (corrected / changed). ² Correction rate among all initially incorrect responses (corrected / wrong).

In clustered logistic regression restricted to initially incorrect responses, GPT-5.0 showed substantially higher odds of correcting errors compared with GPT-3.5 (OR ≈ 5.4, p < 0.001), while GPT-4.0 showed a modest but statistically significant improvement (OR ≈ 2.2, p = 0.046). Prompt language was not independently associated with correction probability after revision (p = 0.33), consistent with the language-parity finding for initial accuracy. The differential correction capacity across models is illustrated in Fig. 2 and Fig. 3.

DISCUSSION

Main Findings and Interpretation

The principal finding of this study is that GPT-5.0 demonstrated significantly higher accuracy on neurosurgery-related multiple-choice questions than its two predecessors, achieving 78.0% overall compared with approximately 61% for both GPT-3.5 and GPT-4.0. The absence of a meaningful

difference between GPT-3.5 and GPT-4.0 ($p = 0.674$ in adjusted analysis) is a notable secondary finding that warrants discussion. One plausible explanation is that GPT-4.0's broader but less curated training base may introduce inconsistencies in highly specialized domains such as neurosurgery, potentially offsetting raw capability gains. This interpretation is consistent with observations that model scale alone does not guarantee improvements in specialized technical fields [14], and that the quality and specificity of training data may be as important as volume [6]. The marked leap observed in GPT-5.0 likely reflects not only expanded data but architectural advances in reasoning and self-consistency [11, 12].

The Level 1 Paradox

A counterintuitive finding requiring explicit discussion is the low accuracy observed at Level 1 for GPT-3.5 and GPT-4.0 (37–38% combined), substantially below performance at Levels 2 through 5. This pattern may reflect several explanations. First, Level 1 in this study targeted basic anatomical and physiological principles rather than applied clinical content. LLMs may process foundational science questions with less contextual scaffolding than clinical scenario-based items, where the clinical framing itself constrains the correct answer. Second, multiple-choice questions in neurosurgery often involve precise anatomical nomenclature, structural relationships, or physiological constants that require exact recall rather than reasoning—a task at which LLMs are known to underperform compared with structured reasoning tasks [7]. Third, the absence of formal psychometric validation of our difficulty classification means that observed level assignments may not map linearly onto actual LLM processing difficulty. This finding underscores the need for empirically derived difficulty metrics—such as item response theory calibration—in future benchmarking studies of this type.

Language Parity

The absence of a statistically significant language effect across all three models is a clinically meaningful finding, particularly for non-English-speaking educational contexts such as Brazil. Despite the well-established predominance of English in LLM training data [6], our results suggest that for neurosurgical content at this complexity range, Portuguese queries yield equivalent performance. GPT-5.0's identical accuracy in both languages (78.0%) is especially noteworthy, suggesting that multilingual generalization has improved in the latest model generation. These findings are broadly consistent with emerging evidence that newer GPT models show reduced language-dependent performance variability [11].

Self-Correction Capacity

The self-correction analysis revealed a clear hierarchy. GPT-5.0 corrected 23.6% and 14.3% of its initial errors in Portuguese and English respectively when prompted with a simple revision cue, compared with only 5.5% and 4.0% for GPT-3.5. This pattern suggests that advanced reasoning architectures confer not only higher initial accuracy but also greater capacity for autonomous error recognition when prompted. This feature may be particularly valuable in educational feedback contexts, where AI can function as an adaptive tutor capable of revising incorrect responses without additional external information.

It should be noted that the revision prompt used in this study (“This is not the correct answer, try again”) provides no diagnostic information beyond the fact of error. A model’s ability to generate a different and correct response under this condition reflects a form of epistemic flexibility—the capacity to hold multiple candidate responses and re-rank them when the initial selection is negated. This is distinct from rote self-correction guided by explicit error feedback, and its increase from GPT-3.5 to GPT-5.0 suggests genuine architectural improvement in this domain.

Limitations

Several limitations must be considered in interpreting these findings.

First, the single-trial design precludes assessment of intra-model variability. LLMs are stochastic systems; querying the same item multiple times may yield different responses, even at default temperature settings. Future studies should incorporate repeated trials per item to estimate response consistency and construct confidence intervals around individual item accuracy.

Second, the question bank was not derived from a formally validated or publicly available standardized examination (e.g., neurosurgery board examination question pools). While items were curated by two neurosurgeons, the inter-rater agreement was not formally quantified (e.g., Cohen’s kappa was not calculated), and the two-evaluator panel is smaller than typically recommended for item bank validation. This limits the generalizability of the reference standard.

Third, the difficulty classification was assigned a priori by clinical judgment rather than by empirical methods such as item response theory. The observed paradoxical performance at Level 1 illustrates the limitations of subjective difficulty stratification.

Fourth, no human reference performance was obtained for comparison. Without data on how neurosurgery residents or specialists perform on the same items, it is not possible to contextualize whether 78% accuracy represents a clinically meaningful threshold. Future studies should include human comparator groups.

Fifth, the self-correction prompt was non-specific and non-informative. While this was a deliberate design choice to isolate autonomous error-correction capacity, it differs substantially from interactive clinical or educational feedback scenarios. More naturalistic prompting paradigms (e.g., chain-of-thought, iterative dialogue) may yield different correction profiles.

Finally, all questions were multiple-choice with four alternatives. This format constrains the type of knowledge assessed, excluding open-ended reasoning, procedural knowledge, and ethical judgment. Performance on closed-format items may not predict performance in real clinical decision support tasks.

CONCLUSION

Among the three GPT versions evaluated, GPT-5.0 demonstrated significantly higher accuracy on neurosurgery-related multiple-choice questions and markedly superior self-correction capacity following a standardized revision prompt. Performance was equivalent in Portuguese and English across all three models, a finding with practical relevance for non-English-speaking medical education environments. Notably, GPT-3.5 and GPT-4.0 did not differ significantly in accuracy on this specialized domain, suggesting that model generation alone does not guarantee performance gains in highly technical fields.

Declarations

Conflict of interest:

The authors have no relevant financial or non-financial interests to disclose.

Human Ethics and Consent to Participate declarations

Not applicable (no human participants)

Statement of AI Use

Artificial intelligence tools were used in the preparation of this manuscript solely for English language revision and editing. Specifically, Claude (Anthropic) and ChatGPT-5 (OpenAI) were employed to improve grammar, syntax, clarity, and overall language quality. No AI tool was used for data collection, statistical analysis, interpretation of results, or generation of scientific content. All intellectual contributions, conclusions, and final editorial decisions remain exclusively the responsibility of the authors.

Funding:

No funding was received to assist with the preparation of this manuscript.

Author Contribution

Conceptualization and supervision: R.L.O.A. Material preparation and question bank curation: R.L.O.A., L.H.C.G., G.S.C., F.C.G. Data collection and accuracy classification: L.H.C.G., G.S.C., P.G.M.E., I.C.L.M.L., R.L.O.A. Statistical analysis: L.H.C.G., R.L.O.A. Writing – original draft: L.H.C.G. Writing – review and editing: R.L.O.A., F.C.G. Supplementary materials: F.C.G., L.H.C.G. All authors read and approved the final manuscript.

Data Availability

All data supporting the findings of this study are available within the paper and furthermore, answer key, prompts, revision request template and the follow-up prompts are available in Supplementary Information including tables and graphics from which we've obtained our data and main results.

References

1. Secinaro S, Calandra D, Secinaro A et al (2021) The role of artificial intelligence in healthcare: a structured review of the literature. *BMC Med Inf Decis Mak* 21:125. <https://doi.org/10.1186/s12911-021-01488-9>
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat Med* 28:31–38. <https://doi.org/10.1038/s41591-021-01614-0>
3. Sharma M, Savage C, Nair M et al (2022) Artificial intelligence and the future of global health. *Lancet Digit Health* 24:e40238. [https://doi.org/10.1016/S2589-7500\(22\)00113-7](https://doi.org/10.1016/S2589-7500(22)00113-7)
4. Langlotz CP (2019) Will artificial intelligence replace radiologists? *Radiol Artif Intell* 1:e190058. <https://doi.org/10.1148/ryai.2019190058>
5. Park KS, Choi H (2024) How to harness the power of GPT for scientific research: a comprehensive review of methodologies, applications, and ethical considerations. *Nucl Med Mol Imaging* 58:323–331. <https://doi.org/10.1007/s13139-024-00876-z>
6. Laizure SC (2024) Caution: ChatGPT does not know what you are asking and does not know what it is saying. *J Pediatr Pharmacol Ther* 29:558–560. <https://doi.org/10.5863/1551-6776-29.5.558>
7. Calabrese G et al (2025) Unraveling the effectiveness of ChatGPT 4.0 for answering common questions from patients undergoing gastrointestinal endoscopy. <https://doi.org/10.1016/j.igie.2025.01.012>. *iGIE*
8. Tang L, Sun Z, Ilday B et al (2023) Evaluating large language models on medical evidence summarization. *NPJ Digit Med* 6:158. <https://doi.org/10.1038/s41746-023-00896-7>
9. Cuthbert R, Simpson A (2023) Artificial intelligence in orthopedics: can ChatGPT pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? *Postgrad Med J* 99:1110–1114. <https://doi.org/10.1093/postmj/qgad053>
10. Daungsupawong H, Wiwanitkit V (2024) Probing artificial intelligence in neurosurgical training: correspondence. *Brain Spine* 4:102751. <https://doi.org/10.1016/j.bas.2024.102751>
11. Choi W, Chang C (2025) ChatGPT-5.0 in education: new capabilities and opportunities for teaching and learning. Preprints. <https://doi.org/10.20944/preprints202508.0684.v1> [Preprint; not peer-reviewed]
12. Leon M (2026) GPT-5 and open-weight large language models: advances in reasoning, transparency, and control. *Inf Syst* 136:102620. <https://doi.org/10.1016/j.is.2025.102620>
13. Handler R et al (2025) The fragile intelligence of GPT-5 in medicine. *Nat Med* 31:3968–3970. <https://doi.org/10.1038/s41591-025-04008-8>

14. Noda R, Tanabe K, Ichikawa D et al (2025) GPT-4’s performance in supporting physician decision-making in nephrology multiple-choice questions. Sci Rep 15:15439. <https://doi.org/10.1038/s41598-025-99774-3>

15. Wang S et al (2025) Capabilities of GPT-5 on multimodal medical reasoning. arXiv preprint. <https://doi.org/10.48550/arXiv.2508.08224>

16. Galiana LI et al (2024) Ethics and artificial intelligence. Rev Clin Esp (Barc) 224:178–186. <https://doi.org/10.1016/j.rceng.2024.02.003>

Figures

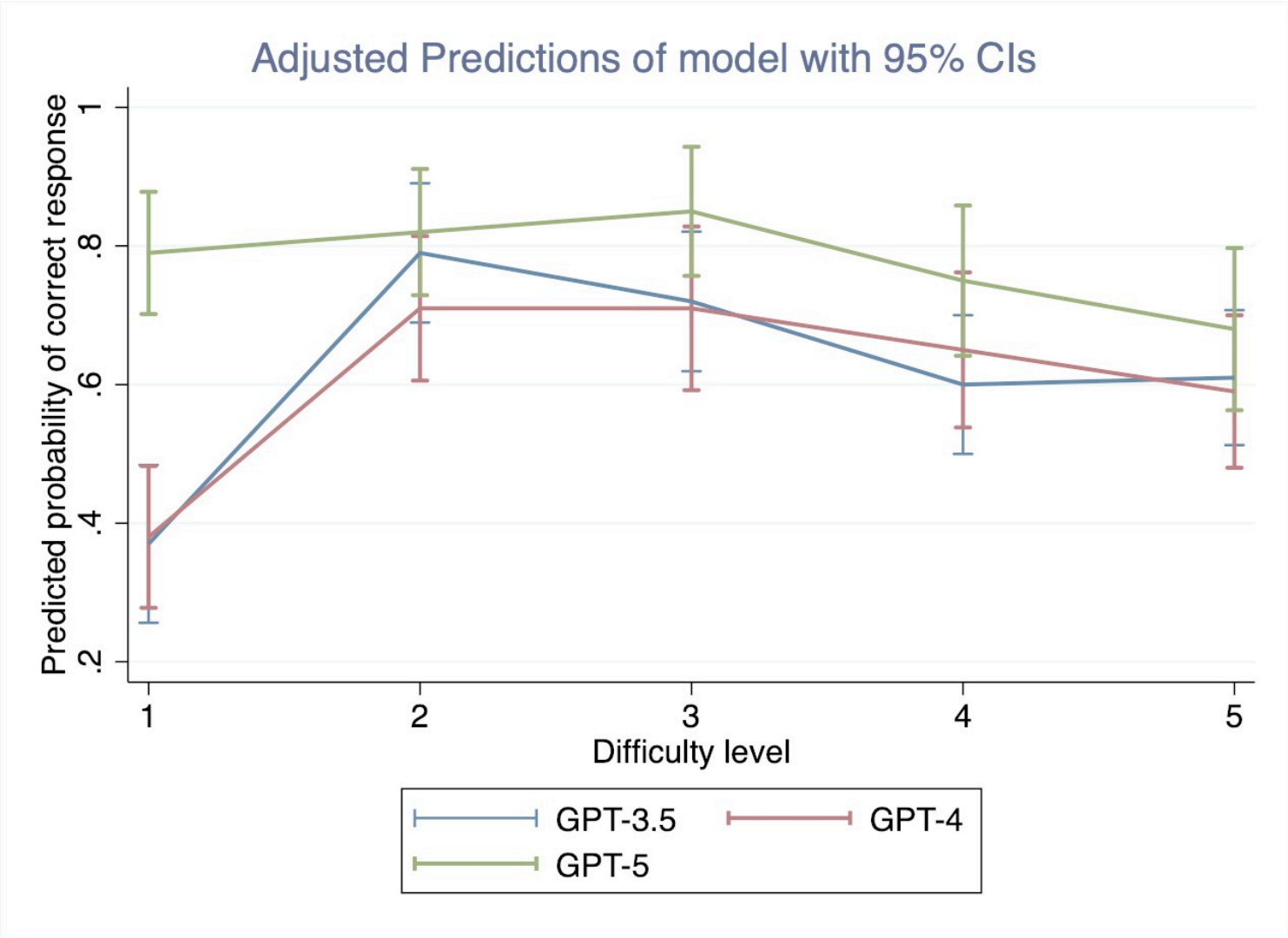


Figure 1

Adjusted predicted probabilities of correct initial responses by difficulty level and model. Predictions were obtained from a clustered logistic regression with a model-by-difficulty interaction, and 95% confidence intervals were reported. GPT-5 showed higher accuracy across all levels, with diminishing relative advantage at higher difficulty.



Figure 2

Odds ratios for correction of initially incorrect responses after a revision request. Forest plot derived from a clustered logistic regression (clustered by item), with GPT 3.5 as the reference. Error bars indicate 95% confidence intervals. GPT 5.0 shows a markedly higher likelihood of correcting errors after revision, while GPT 4.0 shows a modest improvement

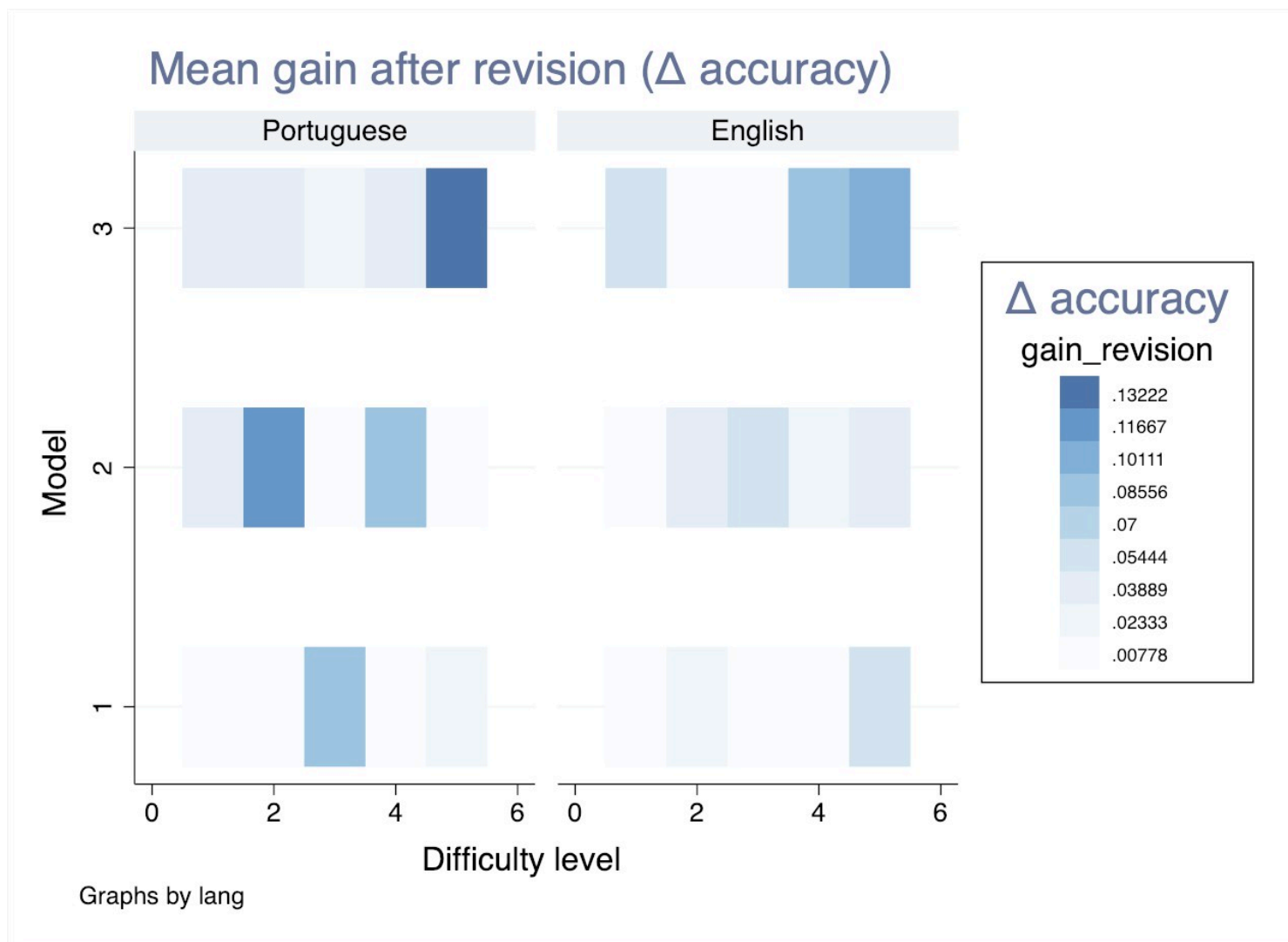


Figure 3

Heatmap of mean accuracy gain after revision (Δ accuracy = final – initial) by model, difficulty level, and language. Darker colors indicate greater improvement following revision. GPT 5.0 shows the largest gains across difficulty levels, with minimal differences between Portuguese and English prompts.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryAppendix.docx](#)