Original Paper

Error Direction of Large Language Models on Medical Multiple-Response Questions: Opportunity-Standardized Benchmark Study

Xinyan Ran[1*], MPH; Haolin Yin[1*], MMed; Yueming Yang[1], MSc; Shanshan Duan[2,3], MD; Xia Luo[1], MBBS; Yang Liu[1*], MMed; Mengqian Liao[1*], MMed

[1]School of Smart Health and Wellness, Chongqing Wuyi Polytechnic, Chongqing, China
[2]School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[3]Beijing Hua Xin Hospital, Beijing, China
*these authors contributed equally

Corresponding Author:

Mengqian Liao, MMed
School of Smart Health and Wellness, Chongqing Wuyi Polytechnic
No. 1001, Yunan Avenue, Banan District
Chongqing 401320
China
Phone: 1 17738378171
Email: liaomq22@tsinghua.org.cn

Abstract

Background: Large language models (LLMs) attain high accuracy on medical examinations, yet evaluations rely predominantly on single-best-answer multiple-choice questions and report only overall accuracy. Multiple-response items, termed X-type in China, require selecting every correct option and remain understudied. Whether models err by omitting correct options or by commissioning incorrect ones has yet to be assessed within an opportunity-standardized framework.

Objective: This study aimed to quantify the accuracy penalty of multiple-response (X-type) items relative to single-select items and to determine, after standardizing for the unequal numbers of correct and incorrect options, whether LLMs lean toward omission or commission.

Methods: We analyzed 1644 text-based items from 10 examination cohorts (2017-2026) of the Chinese National Postgraduate Entrance Examination in Clinical Medicine, comprising 1344 single-select items of types A1, A2, A3, and B, and 300 X-type items. Five LLMs, DeepSeek V4 Pro, Kimi K3, Doubao Seed-2.1-pro, GPT-5.6 Sol, and Gemini-3.6 Flash, were tested zero-shot bilingually, yielding 16,440 planned responses. Item-type effects were estimated using generalized estimating equations with cluster-robust standard errors, adjusting for model, language, discipline, and cohort. Opportunity-standardized error direction was analyzed 2 ways: an option-level model treating each option as 1 judgment record and the omission-bias index.

Results: Of 16,440 planned responses, 16,426 (99.91%) were evaluable. X-type accuracy (2527/3000, 84.23%) was markedly below single-select accuracy (12,990/13,426, 96.75%; adjusted odds ratio at GPT-5.6 Sol in Chinese 0.17, 95% CI 0.11-0.26; $P<.001$), a gap of 12.52 percentage points observed and 13.55 adjusted, with no statistically significant item type × model or item type × language interaction ($P=.19$ and $P=.82$). Raw counts suggested more strict omissions (245/3000, 8.17%) than overselections (201/3000, 6.70%), but X-type items averaged 2.79 correct vs 1.21 incorrect options per item, a 2.3-fold imbalance in error opportunities. After opportunity standardization, the per-opportunity false-selection rate was 6.69% (242/3620), roughly twice the 3.57% per-opportunity omission rate (299/8380; rate ratio 1.87, 95% CI 1.28-2.74; adjusted odds ratio at the reference model and language 0.38, 95% CI 0.20-0.70; $P=.002$), so the tilt is modest on the absolute scale and close to 2-fold on the relative scale. Strict omission showed no between-model difference (all Holm-adjusted pairwise comparisons, $P\geq.46$).

Conclusions: LLMs incur a substantial accuracy penalty on medical multiple-response items. The apparent omission dominance in raw counts is an artifact of answer-key opportunity imbalance; after opportunity standardization, models show a commission tilt that is modest on the absolute scale (3.12 percentage points per opportunity) and close to 2-fold on the relative scale, and that holds in every model-language condition and in the answer-key stratum whose opportunity base

is exactly balanced. Medical LLM benchmarks should report accuracy stratified by item type and error direction on an opportunity-standardized basis.

JMIR Med Educ 2026;12:e111146; doi: 10.2196/111146

Keywords: large language models; benchmark study; error direction; omission and commission; opportunity standardization; omission-bias index; bilingual; artificial intelligence

Introduction

Performance of large language models (LLMs) on medical knowledge examinations has improved rapidly. Med-PaLM 2 scored up to 86.5% on MedQA [1], and GPT-4o reached 90.4% on a 750-item United States Medical Licensing Examination–style question set [2]. Some Chinese-developed models have likewise met or exceeded passing thresholds on China's national medical examinations [3,4]. These results are often read as evidence of medical reasoning, yet the vast majority derive from single-best-answer multiple-choice questions (MCQs).

A single-best-answer item requires selecting 1 best option, whereas a multiple-response item requires selecting the complete correct answer set. Multiple-response items are termed X-type items in China. These formats impose different response requirements, and accordingly they differ in the ways in which a response can be wrong. On an X-type item, omitting a correct option and selecting an incorrect option are qualitatively different errors. Omission and commission are distinct observable error types and should be reported separately when interpreting multiple-response performance.

Prior work has generally reported lower accuracy on multiple response than on single-select items [5-7], but this gap has yet to be decomposed into its omission and commission components. Item-type comparisons have largely been confined to variants of single-best-answer items. For example, one prior study compared only the A1, A2, A3/A4, and B1 single-select variants, excluding multiple-response items [8]. Among studies that examined multiple-response items, some treated omissions and commissions either as symmetric penalty units under partial-credit scoring or merged them into a single error category, while others reported multiple-response items as the lowest-accuracy type, with qualitative discussion of possible causes but no decomposition of error direction [9-11].

Classifications of model errors themselves have remained at the level of reasoning rather than reaching the option level [12]. The general-domain SATA (Select All That Apply)–Bench identified 3 systematic decision biases (unselection, speculation, and count bias) but likewise did not separate the structural effect of the answer key from the model's behavioral tendency [13]. It therefore remains an open question whether, when an LLM errs on a multiple-response item, it tends to omit correct options or to include incorrect ones.

There is an inherent methodological trap in answering this question. The answer keys of multiple-response items typically contain more correct than incorrect options. Omission counts will therefore naturally exceed commission counts even if a model were entirely unbiased toward each option. Comparing the 2 raw counts would mistake this structural feature of the answer key for model conservatism. Distinguishing the structural artifact of answer-key opportunity imbalance from genuine behavioral tendency requires dividing each error count by its own opportunity base. In addition, although the value of question-level answer-aggregation strategies such as self-consistency and ensemble refinement has been demonstrated [1,14], multiple-response items require independent judgment of every option, and whether aggregating at the option level improves performance still lacks direct evidence.

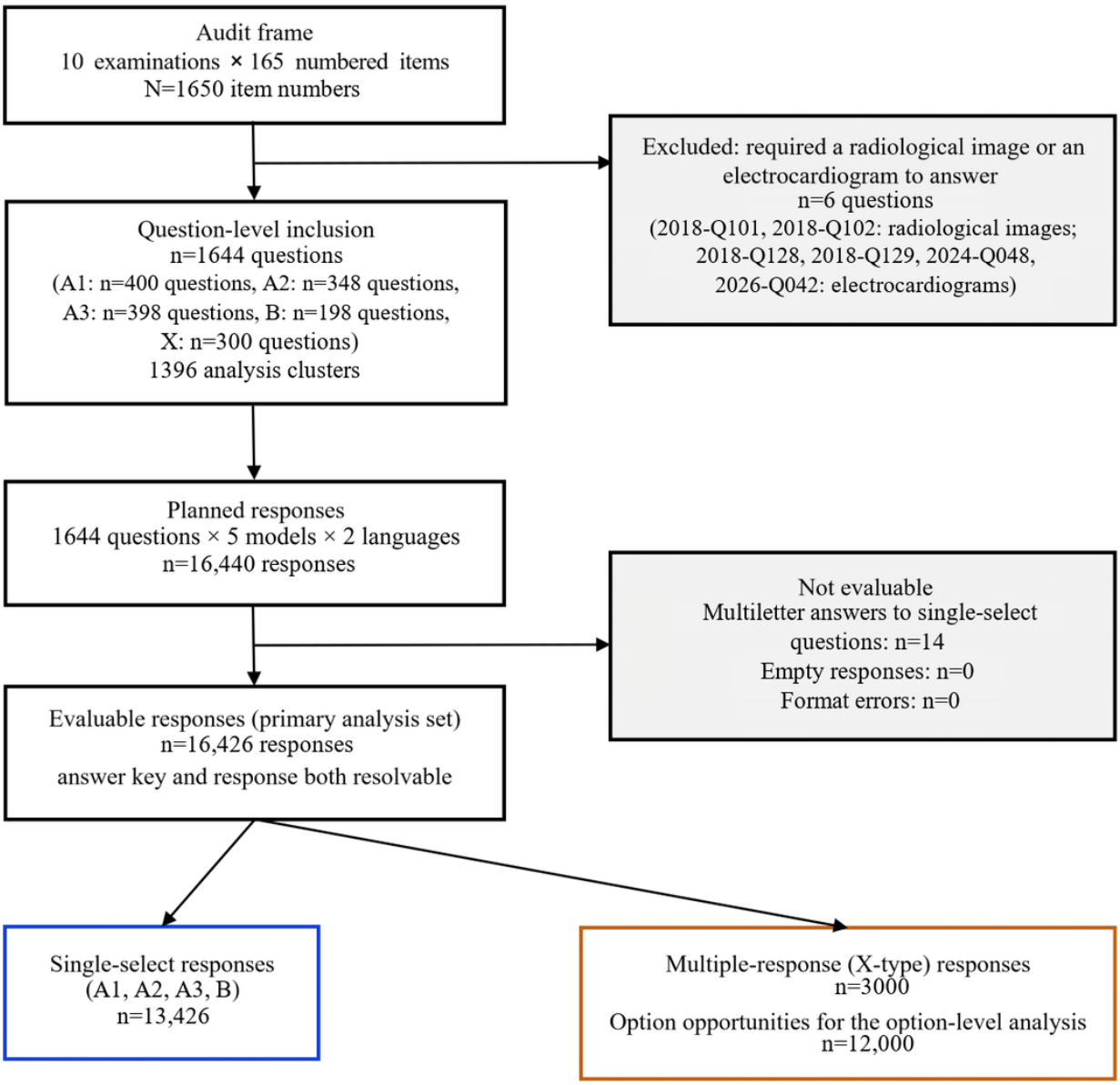
Using 1644 authentic items from 10 consecutive examination cohorts of the Chinese National Postgraduate Entrance Examination in Clinical Medicine, we evaluated 5 mainstream LLMs bilingually. We first quantified the performance penalty of multiple-response items relative to single-select items. We then conducted an opportunity-standardized analysis of error direction through 2 complementary routes: an option-level model and the omission-bias index (OBI). Finally, we tested the effect of option-level majority voting.

Methods

Study Design and Data Source

This cross-sectional benchmarking study used 1644 text-based MCQs from 10 consecutive examination cohorts of the Western Medicine Comprehensive examination, code 306, of the Chinese National Postgraduate Entrance Examination in Clinical Medicine. This is a national examination taken each year by hundreds of thousands of Chinese medical graduates. Each paper contains 165 items, yielding 1650 item numbers across the 10 cohorts, and 6 items that required a radiological image or an electrocardiogram to answer were excluded before analysis. The final sample comprised 400 A1 basic single-select items, 348 A2 stand-alone case-vignette items, 398 A3 shared-stem case items nested within 150 case groups, 198 B matching items, and 300 multiple-response items as shown in Figure 1. Sample flow and analysis-set counts are given in Tables S1 and S2 in Multimedia Appendix 1.

Figure 1. Flow diagram of study sample selection.



Each item was submitted to every model in both its Chinese original and an English translation, giving 16,440 planned responses, and the unit of analysis was the individual response. A3 items share a stem and were clustered by case group, whereas other item types were clustered by item, yielding 1396 clusters, and all inference used cluster-robust methods at this unit and was paired within it. English translations were produced by a bilingual medical researcher and checked for accuracy by a second bilingual reviewer, with disagreements resolved by consensus. One complete bilingual example of each item type is provided in Multimedia Appendix 2. The A3 case-group structure is shown in Figure S1 and Table S3 in Multimedia Appendix 1, and item-type adjudication records are shown in Tables S4 and S5 in Multimedia Appendix 1.

Models and Testing Conditions

Five mainstream LLMs were evaluated, namely, DeepSeek V4 Pro, Kimi K3, Doubao Seed-2.1-pro, GPT-5.6 Sol, and Gemini-3.6 Flash. All models were accessed through their official web interfaces with the highest available reasoning intensity, with web search, retrieval-augmented generation, and cross-session memory disabled; for models without a native offline switch, an instruction forbidding online retrieval was inserted at the start of every session. Testing was zero-shot, with no worked examples or chain-of-thought prompting, and each item was submitted in a separate session to prevent context carryover. Model details and the evaluation protocol are provided in Table S6 in Multimedia Appendix 1. The full verbatim Chinese and English prompts are provided in Multimedia Appendix 3.

Outcome Definitions

Model outputs were standardized before comparison with the reference key. White space and nonalphabetic characters were removed, letters were uppercased, and option letters were extracted, sorted, and deduplicated. For each evaluable response, where K denotes the answer-key set and S the model-selected set, responses were assigned to 5 mutually exclusive categories. These were fully correct when S equals

K, strict omission when S is a strict subset of K, strict overselection when K is a strict subset of S, completely wrong when the 2 sets are disjoint, and mixed errors with both omission and overselection.

Responses that returned multiple-option letters for a single-select item, for which model intent could not be determined, were flagged as type conflicts. Overall accuracy was reported under 2 conventions, an examination-scoring convention that counts type conflicts as errors consistent with the real examination, and an evaluable-response convention that excludes them from the denominator. The primary analysis used the evaluable-response set, with the examination-scoring convention reported alongside as one prespecified sensitivity analysis.

Set-level error-mode proportions (fully correct, strict omission, strict overselection, completely wrong, and mixed errors) and the broad omission rate incorporating mixed errors were reported as descriptive outcomes. The primary directional estimand was the opportunity-standardized error direction, quantified through 2 prespecified complementary routes: the per-opportunity error rates from the option-level model and the OBI, as defined in the next paragraph. The original analysis plan designated the strict omission rate, the proportion of evaluable X-type responses that omitted at least 1 correct option without overselecting, as the primary directional outcome. Because its frequency depends on the number of correct-option opportunities, strict omission is retained as the outcome of the between-model and between-language comparisons, and the directional interpretation rests on the 2 opportunity-standardized routes.

Error direction was analyzed in 2 complementary ways. First, in the option-level model, each option of an X-type item contributed 1 judgment record, and failing to select a correct option or erroneously selecting an incorrect option each counted as 1 error. Option records within an item were not assumed to be behaviorally independent; clustering accommodates within-item dependence in variance estimation but does not make option choices independent Bernoulli trials. The 3000 evaluable X-type responses provided 8380 correct-option opportunities and 3620 incorrect-option opportunities, or 12,000 option-opportunity records in total, allowing comparison of the error rate per omission opportunity with that per false-selection opportunity. Second, the OBI was the within-response proportion of correct options omitted minus the proportion of incorrect options selected, defined as:

OBI = (number of correct options omitted/total correct options) − (number of incorrect options selected/total incorrect options)

A negative OBI indicates a commission, or false selection, tendency. When an answer key exhausts all options, that is, every option is correct, the commission denominator is zero and the OBI is undefined, so the reported OBI is conditional on responses with at least 1 incorrect option opportunity. Such responses are not error-free: they are the only responses on which false selection is impossible by construction, so excluding them removes items that can produce omission but not commission, and the exclusion is directional rather than neutral for an unconditional reading. We therefore added a bounding analysis that assigns those responses their definitional commission component of zero, which makes the index equal to the proportion of correct options omitted and makes the unconditional mean estimable without imputing any unobserved quantity.

Option-Level Agreement and Majority Voting

As an exploratory analysis, intermodel agreement on strict omission was assessed pairwise across the 10 model-language conditions using raw agreement, Cohen $\varkappa$, and Gwet AC1 (AC1), the latter 2 being chance-corrected coefficients, with AC1 reported additionally because $\varkappa$ is sensitive to prevalence and can be markedly depressed when raw agreement is high [15,16]. For the ensemble analysis, option-level majority voting was applied to the 5 models' Chinese responses on the 300 X-type items, with an option counted as selected when at least 3 of the 5 models selected it, a standard combination rule for classifier ensembles [17]. Voting accuracy was compared with that of the best single model; leave-one-model-out voting was used to assess robustness to any single model, and an item-level oracle, which counts an item as correct if at least 1 of the 5 models returned the complete correct answer set, was computed to quantify error complementarity across models. To assess whether the 2 rules differ, voting was further compared with the best single model by pairing on the item and reporting the discordant counts, an exact McNemar test, and a cluster-bootstrap interval for the paired difference, because 2 independent intervals computed on the same items cannot answer that question. The vote threshold was additionally swept from 1 to 5 models, with no threshold selected on the basis of an observed result.

Statistical Analysis

We used generalized estimating equations (GEEs) with a binomial distribution, a logit link, an exchangeable working correlation matrix, and cluster-robust standard errors. The item-type model included model, language, discipline, and cohort and tested item type × model and item type × language interactions; the core analysis used strict omission as the outcome with model, language, and their interaction; and the option-level analysis used option-level error as the outcome with opportunity type and its interactions with model and language. Odds ratios (ORs) and their 95% CIs were obtained from the GEE coefficient covariance. In models with interactions, the reported main-effect OR refers to the reference model and language. Adjusted marginal probabilities were obtained by G-computation, averaging counterfactual predictions over the empirical covariate distribution; their intervals and those of marginal differences used the delta method, with differences retaining the covariance between the 2 marginal estimates (normal-approximation intervals on the probability scale). Interaction terms were assessed with joint Wald tests. Between-model pairwise comparisons used Holm correction, and language effects used within-item exact McNemar tests. Proportions are reported as n/N with Wilson score 95% CIs, and exact *P* values are given. Prespecified sensitivity analyses included

Chinese only, leave-one-cohort-out, strict vs broad omission, the 2 scoring conventions, case-group vs item-level clustering, leave-one-model-out, cluster-bootstrap vs model-based intervals, restriction to items that offered incorrect-option opportunities, exclusion of cohorts whose type labels required adjudication, and exclusion of shared-stem items. The full plan is given in Table S1 in Multimedia Appendix 4, and regression diagnostics are given in Table S2 in Multimedia Appendix 4.

Four analyses not included in the original analysis plan are reported as exploratory without multiplicity correction across analyses: the bounding analysis for responses whose OBI is undefined, stratification of per-opportunity rates by answer-key size, the vote-threshold sweep, and a cohort-recency diagnostic comparing the 2026 cohort with the 2017-2025 cohorts under a recency-based rule that was fixed before any result was inspected.

To test whether the error direction depended on answer-key size, we stratified the per-opportunity analysis by the number of correct options in the key and added answer-key size and its interaction with opportunity type to the option-level model. As sensitivity analyses, all core models were refitted with an independence working correlation matrix and cluster-robust standard errors, and cluster bootstrap intervals (2000 resamples of whole item clusters) were computed for the 2 core per-opportunity rates and their difference. Analyses used Python (version 3.11.5; Python Software Foundation) with the pandas, NumPy, SciPy, and statsmodels packages. Two-sided $P<.05$ was considered statistically significant.

Ethical Considerations

This study used only publicly released examination questions. It involved no human participants, no personal health data, and no animal experiments, so institutional review board approval was not required.

Results

Sample and Data Quality

Dataset characteristics by item type and examination cohort (2017-2026) are shown in Table 1. All 16,440 planned responses were returned with no empty responses, and 14 returned multiple option letters for single-select items and were classified as type conflicts, counted as errors under the examination-scoring convention and excluded under the evaluable-response convention. The final analysis set contained 16,426 evaluable responses, including 3000 X-type responses as shown in Figure 1. Overall accuracy under the 2 conventions was 94.39% and 94.47%, a negligible difference indicating that the choice of convention does not affect the overall conclusions. Response status by model-language condition is given in Table S1 in Multimedia Appendix 5, and item- and response-level exclusions are given in Table S2 in Multimedia Appendix 5.

Table 1. Dataset characteristics by item type and examination cohort, 2017-2026[a].

Cohort	Items included	A1	A2	A3 items	B	X	Analysis clusters	Excluded	Planned responses	Evaluable responses	Type conflicts
2017	165	40	35	40	20	30	140	0	1650	1650	0
2018	161	40	35	38	18	30	138	4	1610	1606	4
2019	165	40	35	40	20	30	140	0	1650	1650	0
2020	165	40	35	40	20	30	140	0	1650	1650	0
2021	165	40	35	40	20	30	140	0	1650	1650	0
2022	165	40	35	40	20	30	140	0	1650	1648	2
2023	165	40	35	40	20	30	140	0	1650	1649	1
2024	164	40	34	40	20	30	139	1	1640	1639	1
2025	165	40	35	40	20	30	140	0	1650	1645	5
2026	164	40	34	40	20	30	139	1	1640	1639	1
Total	1644	400	348	398	198	300	1396	6	16,440	16,426	14

[a]Of 1650 item numbers, 6 image-dependent items were excluded. Planned responses equal 1644 items × 5 models × 2 languages, or 16,440. Empty and format-error responses were 0 across all cohorts and are omitted for brevity; the full response-status table is provided in Table S1 in Multimedia Appendix 5.

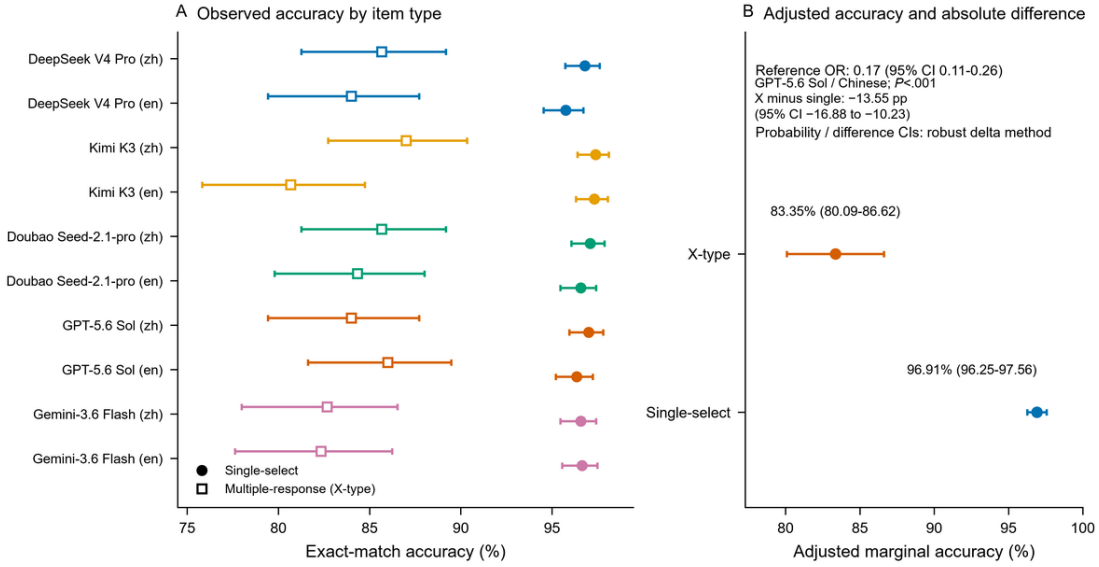
Item-Type Penalty

Pooling models and languages, overall accuracy was 94.47% (15,517/16,426; 95% CI 94.11%-94.81%), with 96.75% (12,990/13,426; 95% CI 96.44%-97.04%) for single-select items and 84.23% (2527/3000; 95% CI 82.89%-85.49%) for X-type items. Among evaluable X-type responses, 245 of 3000 (8.17%; 95% CI 7.24%-9.20%) were strict omissions, 272 of 3000 (9.07%; 95% CI 8.09%-10.15%) were broad omissions, and 201 of 3000 (6.70%; 95% CI 5.86%-7.65%) were strict overselections.

After adjustment for model, language, discipline, and cohort, X-type accuracy was significantly lower than single-select accuracy (reference-condition OR at GPT-5.6 Sol in Chinese 0.17, 95% CI 0.11-0.26; $P<.001$; 16,426 responses, 1396 clusters). The covariate-adjusted marginal difference (X-type minus single-select) was −13.55 percentage points (95% CI −16.88 to −10.23), with adjusted marginal

accuracy falling from 96.91% (95% CI 96.25%-97.56%) for single-select to 83.35% (95% CI 80.09%-86.62%) for X-type, compared with an observed difference of −12.52 percentage points, as shown in Figure 2A and B. The item type × model interaction was nonsignificant (Wald χ^2_4=6.13, *P*=.19), as was the item type × language interaction (Wald χ^2_1=0.05, *P*=.82), providing no statistically significant evidence of heterogeneity, without establishing equivalent effects across models or languages. The estimate was unchanged after excluding shared-stem items (OR 0.17, 95% CI 0.11-0.24), showing that the item-type association persisted in this sensitivity analysis; it does not isolate a causal effect of the response requirement. Accuracy by analysis item type is shown in Figure S1 and Table S3 in Multimedia Appendix 5.

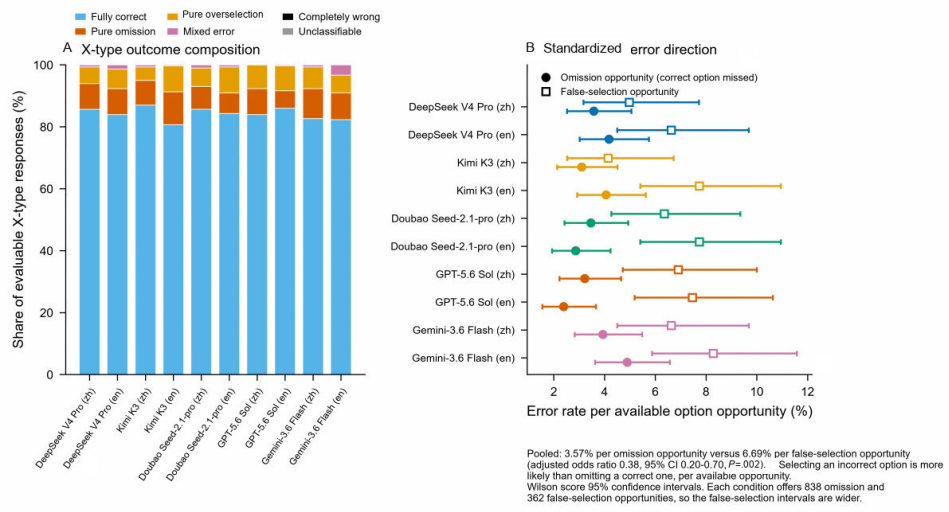
Figure 2. Accuracy by item type. (A) Observed overall accuracy for each model-language condition, with single-select items as filled circles and X-type items as open squares and Wilson 95% CIs. (B) Adjusted marginal accuracy and the X-type-minus-single-select difference, standardized over the fixed empirical covariate distribution after adjustment for model, language, discipline, and examination cohort across 1396 clusters. Intervals in (B) are cluster-robust delta-method 95% CIs. The odds ratio (OR) and its robust CI refer to GPT-5.6 Sol in Chinese.



Error Direction After Opportunity Standardization

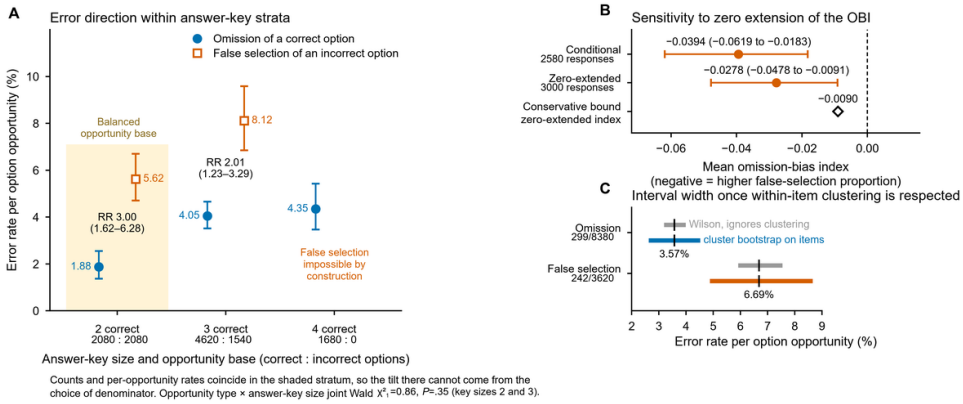
Among evaluable X-type responses, 245 of 3000 (8.17%; 95% CI 7.24%-9.20%) strict omissions and 201 of 3000 (6.70%; 95% CI 5.86%-7.65%) strict overselections superficially suggested more omissions than commissions. However, the 3000 evaluable X-type responses provided 8380 correct options with a mean of 2.79 per item and 3620 incorrect options with a mean of 1.21 per item, so omission opportunities outnumbered false-selection opportunities 2.3-fold. After standardizing by the number of opportunities for each error, the results reversed. A total of 299 omissions arose from 8380 omission opportunities (3.57%; 95% CI 3.19%-3.99%), whereas 242 false selections arose from 3620 false-selection opportunities (6.69%; 95% CI 5.92%-7.55%), and the error rate per false-selection opportunity was roughly twice that per omission opportunity (rate ratio 1.87, 95% CI 1.28-2.74; adjusted OR 0.38, 95% CI 0.20-0.70; *P*=.002; 12,000 option opportunities, 300 clusters) as shown in Figure 3A and B. The OBI provided independent verification. Pooled across conditions it was –0.039 with 95% CI –0.062 to –0.018 based on 2580 contributing responses, while 420 responses were undefined because the key exhausted all options. Response-level distribution is shown in Figure S1 in Multimedia Appendix 6. The strict omission rate rose with the number of correct options as shown in Figure S2 in Multimedia Appendix 6, consistent with an opportunity-structure explanation. Error-mode composition by condition is shown in Table S1 in Multimedia Appendix 6, option-level error rates in Table S2 in Multimedia Appendix 6, and answer-key and selected-set size distributions in Tables S3 and S4 in Multimedia Appendix 6. Together these results show that the raw omission dominance of 8.17% vs 6.70% was driven by the larger number of correct options rather than model conservatism, and after controlling for opportunity counts models instead leaned toward commission.

Figure 3. Error modes and error direction. (A) Composition of mutually exclusive outcome categories among evaluable X-type responses. (B) Per-option-opportunity error rate by condition (838 omission opportunities and 362 false-selection opportunities per model-language condition), with omission opportunities as filled circles and false-selection opportunities as open squares. Error bars are Wilson 95% CIs, which are descriptive and do not correct for within-item clustering; annotations give the pooled rates and the adjusted OR of omission relative to false selection; and cluster bootstrap 95% CIs for the pooled rates are reported in the Results section.



The index was undefined for 420 responses, not because those responses were error-free but because the key exhausted all options, leaving a zero commission denominator. These are the only responses on which false selection is impossible by construction (Figure 4A). Assigning them their definitional commission component of zero gives a mean omitted-option proportion of 0.0435 (95% CI 0.0208-0.0690, cluster bootstrap) in that subgroup and an unconditional mean OBI across all 3000 responses of –0.028 (95% CI –0.048 to –0.009), still negative. The direction also holds without using the observed value. Attributing all 299 option-level omissions to the 1680 correct-option opportunities contributed by these responses bounds the unconditional mean at –0.009, and sign reversal would require an omitted-option proportion above 0.242, which the data cannot produce, as shown in Figure 4B.

Figure 4. Error direction robustness. (A) Per-opportunity error rates stratified by answer-key size (2 or 3 correct options); the shaded band marks the balanced-opportunity baseline; items with 4 correct options permit omission but no false selection by construction. (B) Mean OBI conditional on a defined index (–0.039), unconditional after assigning definitional zero commission (–0.028), and the worst-case bound (–0.009). (C) Wilson score intervals vs cluster-bootstrap intervals for the pooled per-opportunity rates, showing wider intervals once within-item clustering is respected. OBI: omission-bias index.



Within the stratum of items keyed with 2 correct options, correct and incorrect opportunities were equal at 2080 each, so standardization has no effect; omissions numbered 39 (1.88%) against 117 false selections (5.62%; rate ratio 3.00, 95% CI 1.62-6.28). In items keyed with 3 correct options, the rates were 4.05% (187/4620) vs 8.12% (125/1540; rate ratio 2.01, 95% CI 1.23-3.29). Items keyed with 4 correct options contain no incorrect options, making false selection structurally impossible; their omission rate was 4.35% (73/1680). The interaction gave no evidence that the direction changed with answer-key size (Wald χ^2_1=0.86, P=.35), and the false-selection-to-omission rate ratio was largest in the balanced stratum, whereas the absolute rate difference was slightly larger in the 3-correct-option stratum (Figure 4A and Table 2).

Table 2. Adjusted regression results for the item-type effect, between-model strict omission, and option-level error direction, with answer-key-size strata[a].

Analysis and contrast	Effect measure	Estimate	Adjusted absolute difference	*P* value	Holm-adjusted *P* value	Observations	Clusters
X-type vs single-select accuracy							
X-type vs single-select (reference model and language)	Adjusted OR[b] (95% CI)	0.17 (0.11-0.26)	−13.55 percentage points, covariate-adjusted marginal (95% CI −16.88 to −10.23; −12.52 observed)	<.001	NA[c]	16,426	1396
Joint Wald: item type × model interaction	Wald χ^2	6.13 (4)	—[d]	.19	NA	16,426	1396
Joint Wald: item type × language interaction	Wald χ^2	0.05 (1)	—	.82	NA	16,426	1396
Between-model strict omission							
DeepSeek V4 Pro vs GPT-5.6 Sol (reference language zh)	Adjusted OR (95% CI)	1.00 (0.68-1.48)	—	>.99	NA	3000	300
Kimi K3 vs GPT-5.6 Sol (reference language zh)	Adjusted OR (95% CI)	0.96 (0.61-1.49)	—	.84	NA	3000	300
Doubao Seed-2.1-pro vs GPT-5.6 Sol (reference language zh)	Adjusted OR (95% CI)	0.87 (0.53-1.43)	—	.58	NA	3000	300
Gemini-3.6 Flash vs GPT-5.6 Sol (reference language zh)	Adjusted OR (95% CI)	1.18 (0.82-1.71)	—	.37	NA	3000	300
English vs Chinese (reference model)	Adjusted OR (95% CI)	0.65 (0.42-1.02)	—	.06	NA	3000	300
Between-model strict omission (pairwise, family F_1)							
Kimi K3 vs DeepSeek V4 Pro	Adjusted OR (95% CI)	1.14 (0.82-1.58)	1.00 percentage points	.45	>.99	1200	300
Doubao Seed-2.1-pro vs DeepSeek V4 Pro	Adjusted OR (95% CI)	0.82 (0.58-1.17)	−1.33 percentage points	.28	>.99	1200	300
GPT-5.6 Sol vs DeepSeek V4 Pro	Adjusted OR (95% CI)	0.82 (0.60-1.13)	−1.33 percentage points	.23	>.99	1200	300
Gemini-3.6 Flash vs DeepSeek V4 Pro	Adjusted OR (95% CI)	1.11 (0.77-1.61)	0.83 percentage points	.56	>.99	1200	300
Doubao Seed-2.1-pro vs Kimi K3	Adjusted OR (95% CI)	0.73 (0.53-0.99)	−2.33 percentage points	.046	.46	1200	300
GPT-5.6 Sol vs Kimi K3	Adjusted OR (95% CI)	0.72 (0.51-1.03)	−2.33 percentage points	.07	.63	1200	300
Gemini-3.6 Flash vs Kimi K3	Adjusted OR (95% CI)	0.98 (0.69-1.39)	−0.17 percentage points	.91	>.99	1200	300
GPT-5.6 Sol vs Doubao Seed-2.1-pro	Adjusted OR (95% CI)	1.00 (0.65-1.53)	0.00 percentage points	>.99	>.99	1200	300
Gemini-3.6 Flash vs Doubao Seed-2.1-pro	Adjusted OR (95% CI)	1.35 (0.93-1.96)	2.17 percentage points	.11	.77	1200	300
Gemini-3.6 Flash vs GPT-5.6 Sol	Adjusted OR (95% CI)	1.36 (0.97-1.91)	2.17 percentage points	.08	.63	1200	300
Between-model strict omission							
Joint Wald: model × language interaction	Wald χ^2	6.92 (4)	—	.14	NA	3000	300
Option-level error direction							
Omission opportunity vs false-selection opportunity (reference model and language)	Adjusted OR (95% CI)	0.38 (0.20-0.70)	−3.93 percentage points, average marginal effect (−5.98 to −1.88)	.002	NA	12,000	300
Option-level error direction (exploratory)							
Adjusted marginal probability, omission opportunity	Adjusted probability (95% CI)	3.35% (2.42-4.29)	—	NA	NA	12,000	300
Adjusted marginal probability, false-selection opportunity	Adjusted probability (95% CI)	7.29% (5.38-9.20)	—	NA	NA	12,000	300

(page number not for citation purposes)

Analysis and contrast	Effect measure	Estimate	Adjusted absolute difference	*P* value	Holm-adjusted *P* value	Observations	Clusters
Rate ratio, false selection relative to omission	Rate ratio (95% CI)	1.87 (1.28-2.74)	—	NA	NA	12,000	300
Joint Wald: opportunity type × model	Wald χ^2	4.22 (4)	—	.38	NA	12,000	300
Joint Wald: opportunity type × language	Wald χ^2	1.88 (1)	—	.17	NA	12,000	300
Option-level error direction, by answer-key size (exploratory)							
Answer key with 2 correct options (1040 responses, balanced opportunity base)	Omission vs false selection, n/N (%)	39/2080 (1.88%) vs 117/2080 (5.62%)	3.75 percentage points; rate ratio 3.00 (1.62-6.28)	NA	NA	4160	104
Answer key with 3 correct options (1540 responses)	Omission vs false selection, n/N (%)	187/4620 (4.05%) vs 125/1540 (8.12%)	4.07 percentage points; rate ratio 2.01 (1.23-3.29)	NA	NA	6160	154
Answer key with 4 correct options (420 responses)	Omission vs false selection, n/N (%)	73/1680 (4.35%); false selection impossible	NA	NA	NA	1680	42
Joint Wald: opportunity type × answer-key size	Wald χ^2	0.86 (1)	—	.35	NA	10,320	258

[a]Binomial-logit generalized estimating equations use an exchangeable working correlation and item-cluster-robust standard errors, adjusting for model, language, discipline, and categorical cohort; independence-correlation refits are sensitivity analyses. In interaction models, odds ratios refer to GPT-5.6 Sol in Chinese; pooled contrasts are average marginal differences with delta-method 95% CIs from the cluster-robust covariance. Interactions use joint Wald tests. The 10 pairwise strict-omission comparisons form 1 Holm-adjusted family (F_1). Each X-type option contributes 1 judgment, clustered by item. Rows labeled exploratory, including the answer-key-size strata, were added during revision and are interpreted as exploratory. The 4-correct-option stratum has no false-selection opportunities; the key-size interaction test includes only sizes 2 and 3.

[b]OR: odds ratio.

[c]NA, not applicable.

[d]Not available.

Cross-Model and Cross-Language Consistency

Strict omission rates across the 10 model-language conditions ranged from 5.67% (17/300) for GPT-5.6 Sol in English to 10.67% (32/300) for Kimi K3 in English as shown in Table 3 and Figure 5A and B. With GPT-5.6 Sol in Chinese as reference, adjusted ORs across models spanned 0.87 to 1.18 with all confidence intervals including 1 (3000 responses, 300 clusters; Table 2), and no pairwise comparison survived Holm correction, with the smallest adjusted *P* value equal to .46. The model-language interaction was nonsignificant (Wald χ^2_4=6.92, *P*=.14), and the adjusted OR for English vs Chinese was 0.65 (95% CI 0.42-1.02; *P*=.06). Within-item paired analysis showed between-language accuracy differences of –0.06 to +1.22 percentage points and strict-omission differences of –2.67 to +2.67 percentage points across models, none significant after Holm correction, with the smallest adjusted *P* value equal to .22 and .48, respectively, as shown in Figure 6A and B. Accuracy and omission rates by cohort and discipline are shown in Figure S2 and Tables S4 and S5 in Multimedia Appendix 5, paired details in Tables S5 and S6 in Multimedia Appendix 6, and discipline item counts in Table S6 in Multimedia Appendix 5.

Table 3. Performance of the 5 large language models in Chinese and English by item type, abridged[a].

Model	Language	Overall, n/N (%)	Single-select, n/N (%)	X-type, n/N (%)	Strict omission, n/N (%)	Strict overselection, n/N (%)	Mean OBI[b]
DeepSeek V4 Pro	Chinese	1557/1643 (94.8)	1300/1343 (96.8)	257/300 (85.7)	25/300 (8.3)	16/300 (5.3)	–0.026
DeepSeek V4 Pro	English	1538/1643 (93.6)	1286/1343 (95.8)	252/300 (84)	25/300 (8.3)	19/300 (6.3)	–0.035
Kimi K3	Chinese	1569/1643 (95.5)	1308/1343 (97.4)	261/300 (87)	24/300 (8)	13/300 (4.3)	–0.008
Kimi K3	English	1550/1644 (94.3)	1308/1344 (97.3)	242/300 (80.7)	32/300 (10.7)	25/300 (8.3)	–0.045
Doubao Seed-2.1-pro	Chinese	1562/1644 (95)	1305/1344 (97.1)	257/300 (85.7)	22/300 (7.3)	18/300 (6)	–0.042
Doubao Seed-2.1-pro	English	1550/1643 (94.3)	1297/1343 (96.6)	253/300 (84.3)	20/300 (6.7)	25/300 (8.3)	–0.060
GPT-5.6 Sol	Chinese	1550/1638 (94.6)	1298/1338 (97)	252/300 (84)	25/300 (8.3)	23/300 (7.7)	–0.037
GPT-5.6 Sol	English	1551/1642 (94.5)	1293/1342 (96.3)	258/300 (86)	17/300 (5.7)	24/300 (8)	–0.056
Gemini-3.6 Flash	Chinese	1544/1642 (94)	1296/1342 (96.6)	248/300 (82.7)	29/300 (9.7)	21/300 (7)	–0.034
Gemini-3.6 Flash	English	1546/1644 (94)	1299/1344 (96.7)	247/300 (82.3)	26/300 (8.7)	17/300 (5.7)	–0.049

[a]Denominators are shown for every proportion. Overall and single-select accuracy use all evaluable responses, while X-type metrics use evaluable X-type responses. The OBI is undefined when the answer key exhausts the option set, leaving a zero commission denominator. This applies to 42 responses per condition, 420 in total, so OBI n=258 per condition and 2580 in total. Those responses are not error-free. They are the only responses on which false selection is impossible by construction. CIs are Wilson score intervals. The full error-pattern breakdown and option-level error rates, with all CIs, are provided in Tables S1 and S2 in Multimedia Appendix 6.
[b]OBI: omission-bias index.

Figure 5. Strict omission by model. (A) Adjusted ORs and 95% CIs for all pairwise between-model comparisons on a log scale, with Holm-adjusted *P* values. (B) Adjusted strict-omission probability by model-language condition, with Chinese as filled and English as open markers, and cluster-robust delta-method 95% CIs for probabilities standardized over the fixed empirical covariate distribution. Gray markers and bars show observed rates and descriptive Wilson 95% CIs. GPT-5.6 Sol in Chinese served as the reference condition.

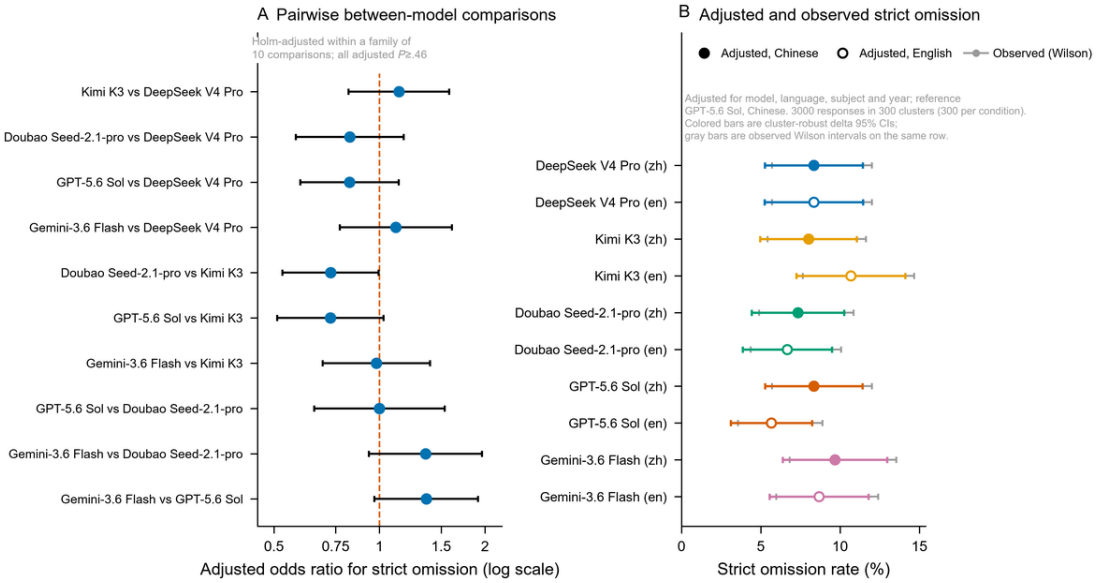
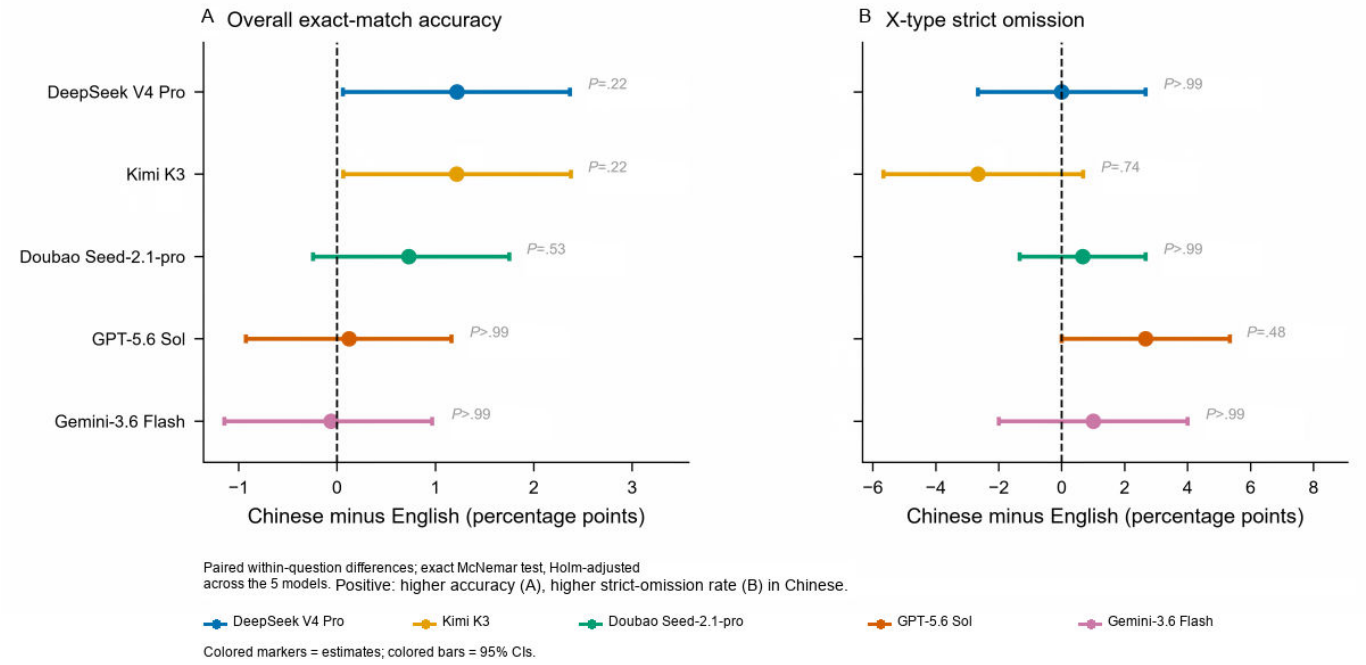


Figure 6. Language effects. Within-item paired differences of Chinese minus English in (A) overall accuracy and (B) X-type strict omission, with cluster-bootstrap 95% CIs. The dashed line marks no difference; positive values indicate higher accuracy (A) and strict-omission rate (B) in Chinese. Holm-adjusted exact McNemar *P* values are labeled for each model.



Option-Level Majority Voting

In an exploratory analysis, raw pairwise agreement between models on strict omission was 90% to 94%, with Cohen κ of 0.33 to 0.59 and AC1 of 0.88 to 0.93 as shown in Figure S3 and Table S7 in Multimedia Appendix 6. On the 300 Chinese X-type items, option-level majority voting, which counts an option when at least 3 of 5 models select it, achieved 89% accuracy (95% CI 84.95%-92.06%), above the best single model at 87% as shown in Figure S4 in Multimedia Appendix 6. Voting accuracy ranged from 86.33% to 87.33% after removing any single model, entirely below the full 5-model accuracy of 89% and reducing its apparent advantage over the best single model to at most 0.33 percentage points. Paired on the item, the 2.00-percentage-point margin of the full 5-model vote rests on 22 discordant items, 14 favoring voting and 8 against. The paired difference is 2.00 percentage points (95% CI −1.33 to 5.33; exact McNemar *P*=.29), which is not distinguishable from zero. Because the comparator was selected as the best of 5 models on these same data, this comparison is conservative for the ensemble. Sweeping the vote threshold from 1 to 5 models showed a monotone trade-off between the 2 error directions. The per-opportunity omission rate rose from 0.72% to 7.76% while the false-selection rate fell from 14.36% to 1.10%. The 3-of-5 rule used here gave 2.74% omission against 3.59% false selection, and no threshold cleared the best single model by a detectable margin (Table 4 and Figure 7A -C). The item-level oracle, which counts an item as correct if at least 1 model returned the complete correct answer set, was 96% (288/300), indicating substantial complementary errors across models. Full voting-strategy results are shown in Table S9 in Multimedia Appendix 6, and exploratory model-origin grouping in Table S8 in Multimedia Appendix 6.

Table 4. Ensemble vote-threshold sweep with paired comparisons against the best single model (exploratory)[a].

Language	Vote threshold	Exact-match accuracy, n/N (%, 95% CI)	Per-opportunity omission, n/N (%)	Per-opportunity false selection, n/N (%)	Paired difference vs best single model, pp[b] (95% CI)	Discordant items (vote only/single only)	Exact McNemar, *P* value	Holm-adjusted, *P* value
Chinese	≥1 of 5 models	247/300 (82.33, 77.62 to 86.23)	6/838 (0.72)	52/362 (14.36)	−4.67 (−9.33 to −0.33)	17/31	.06	.24
Chinese	≥2 of 5 models	264/300 (88.00, 83.83 to 91.20)	14/838 (1.67)	26/362 (7.18)	1.00 (−2.67 to 4.33)	16/13	.71	.86
Chinese	≥3 of 5 models	267/300 (89.00, 84.95 to 92.06)	23/838 (2.74)	13/362 (3.59)	2.00 (−1.33 to 5.33)	14/8	.29	.86
Chinese	≥4 of 5 models	256/300 (85.33, 80.88 to 88.89)	37/838 (4.42)	10/362 (2.76)	−1.67 (−5.00 to 1.67)	10/15	.42	.86
Chinese	≥5 of 5 models	240/300 (80.00, 75.11 to 84.13)	65/838 (7.76)	4/362 (1.10)	−7.00 (−11.00 to −3.00)	9/30	.001	.005
Chinese	Oracle bound (diagnostic only)	288/300 (96.00, 93.14 to 97.70)	—[c]	—	—	—	NA[d]	NA
Chinese	Best single model: Kimi K3	261/300 (87.00, 82.72 to 90.34)	—	—	—	—	NA	NA
English	≥1 of 5 models	236/300 (78.67, 73.68 to 82.92)	8/838 (0.95)	65/362 (17.96)	−7.33 (−11.33 to −3.33)	7/29	<.001	.002
English	≥2 of 5 models	257/300 (85.67, 81.25 to 89.18)	13/838 (1.55)	35/362 (9.67)	−0.33 (−3.00 to 2.33)	8/9	>.99	>.99
English	≥3 of 5 models	259/300 (86.33, 81.98 to 89.76)	24/838 (2.86)	21/362 (5.80)	0.33 (−2.33 to 3.33)	10/9	>.99	>.99
English	≥4 of 5 models	256/300 (85.33, 80.88 to 88.89)	37/838 (4.42)	11/362 (3.04)	−0.67 (−3.67 to 2.67)	12/14	.85	>.99
English	≥5 of 5 models	234/300 (78.00, 72.97 to 82.32)	72/838 (8.59)	5/362 (1.38)	−8.00 (−12.67 to −3.33)	15/39	.001	.006
English	Oracle bound (diagnostic only)	285/300 (95.00, 91.92 to 96.95)	—	—	—	—	NA	NA
English	Best single model: GPT-5.6 Sol	258/300 (86.00, 81.62 to 89.47)	—	—	—	—	NA	NA

[a]At threshold k, the ensemble selects options chosen by at least k of the 5 models; an empty answer set is incorrect. Exact-match accuracy is calculated on the 300 items evaluable by all models per language. Omission and false-selection denominators are 838 correct and 362 incorrect options per language, respectively. Paired differences compare each threshold with the best single model on these items; 95% CIs use an item-cluster bootstrap, *P* values use exact McNemar tests, and Holm adjustment covers 5 thresholds per language. The item-level oracle counts an item as correct if at least 1 of the 5 models returned the complete correct answer set. It requires knowledge of the correct answers and is a diagnostic benchmark, not a deployable strategy.

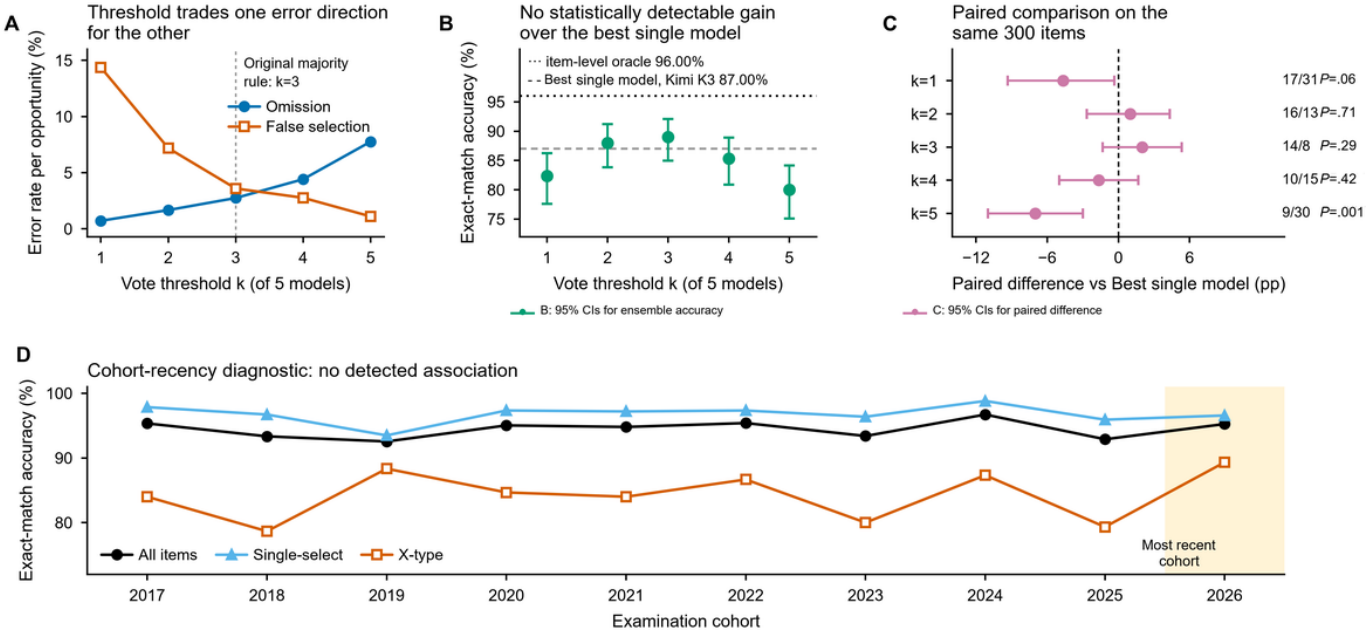
[b]pp: percentage points.

[c]Not available.

[d]NA: not applicable.

(page number not for citation purposes)

Figure 7. Ensemble vote-threshold sweep and cohort-recency diagnostic. (A) Trade-off between omission and false selection across vote thresholds k=1 to 5; the original majority-vote rule uses k=3. (B) Exact-match accuracy by threshold; k=2 and k=3 have numerically higher accuracy than the best single model (Kimi K3, 87%), but no evaluated threshold outperformed it by a statistically detectable margin in paired comparisons. The item-level oracle (96%) is a diagnostic benchmark requiring knowledge of the correct answers. (C) Paired differences in accuracy between voting at each threshold and the best single model. (D) Accuracy by examination cohort for all items, single-select items, and X-type items; the shaded band marks the most recent cohort (2026), with no monotonic trend across the decade. pp: percentage points.



Sensitivity Analyses

The principal findings were stable across the prespecified sensitivity-analysis families. Between-model ORs for Chinese-only responses were 0.87 to 1.18. Under leave-one-cohort-out refitting, model contrasts ranged from 0.86 to 1.61, and some intervals were compatible with modest differences. These sensitivity analyses did not establish a stable model ranking, as shown in Figure S1 in Multimedia Appendix 4. The item-type effect was similar under case-group clustering (OR 0.16, 95% CI 0.11-0.22; 1396 clusters) and item clustering (OR 0.17, 95% CI 0.12-0.23; 1644 clusters). As a consistency check, the OBI remained negative after restriction to items that offered incorrect-option opportunities (–0.039, 95% CI –0.0619 to –0.0183), and cluster-bootstrap and model-based intervals agreed. Results of the prespecified sensitivity-analysis families are shown in Table S1 in Multimedia Appendix 4.

Refitting all core models with an independence working correlation matrix and cluster-robust standard errors left every conclusion unchanged: the item-type OR moved from 0.170 to 0.173, the language contrast OR stayed at 0.654, and the option-level directionality OR moved from 0.38 (95% CI 0.20-0.70; *P*=.002) to 0.43 (95% CI 0.23-0.80; *P*=.008). Cluster bootstrap intervals were wider than Wilson intervals but supported the same conclusion: omission 3.57% (95% CI 2.63%-4.50%; Wilson 3.19-3.99), false selection 6.69% (95% CI 4.84%-8.66%; Wilson 5.92-7.55), and difference –3.12 percentage points (95% CI –5.24 to –1.10) (Figure 4C).

Training-Data Contamination

A cohort-recency diagnostic found no signal in either direction. Exact-match accuracy was 1561/1639 (95.24%) in the 2026 cohort against 13,956/14,787 (94.38%) in the 2017-2025 cohorts, an observed difference of 0.86 percentage points and an adjusted OR of 1.18 (95% CI 0.67-2.06; *P*=.56; 1396 clusters), with no evidence that the contrast differed across models (Wald χ^2_4=2.21, *P*=.70). Treating cohort year as continuous gave no trend across the decade (OR per year 1.011, 95% CI 0.957-1.069; *P*=.69), and annual accuracy varied nonmonotonically from 92.55% in 2019 to 96.71% in 2024. Redefining the recent group as 2025-2026 gave 94.06% vs 94.57% (adjusted OR 0.88, 95% CI 0.60-1.29; *P*=.50). Stratified by item type, single-select accuracy was 96.56% vs 96.77% (adjusted OR 0.85, 95% CI 0.42-1.70; *P*=.65), leaving little room for a detectable difference at that ceiling, whereas X-type accuracy was 89.33% vs 83.67% (adjusted OR 1.58, 95% CI 0.68-3.63; *P*=.29), as shown in Figure 7D. Because cohort year is confounded with paper difficulty, discipline composition, time in public circulation, and unknown training cutoff dates, and because a single cohort contributes only about 1640 responses, this diagnostic is indirect: it can neither establish nor exclude overlap between the item bank and model training corpora.

Discussion

Principal Findings

Using 1644 authentic items across 10 cohorts, this zero-shot, bilingual evaluation of 5 LLMs yielded 3 linked findings. First, multiple-response items carried a large accuracy penalty relative to single-select items, 12.52 percentage points observed and 13.55 percentage points after adjustment, consistent in direction across all 5 models and both languages. Second and more importantly, error direction reversed

after opportunity standardization. Raw counts showed more omissions than commissions, but this was an artifact of the answer-key opportunity imbalance, that is, a 2.3-fold excess of correct over incorrect options. Once standardized by opportunity, the false-selection error rate was roughly twice the omission error rate: a modest 3.12 percentage points on the absolute scale but a rate ratio of 1.87 (95% CI 1.28-2.74) on the relative scale. Third, in an exploratory analysis, option-level majority voting reached 89% accuracy against 87% for the best single model, but paired on the item this margin was not distinguishable from zero (P=.29), and the vote threshold traded omission against false selection; the 96% oracle bound indicates substantial error complementarity across models.

Comparison With Prior Work

Multiple-choice examination scores are a widely used evaluation metric, and as with other education technologies judged against traditional standards, such as virtual reality in anatomy education [18], benchmark results are best interpreted against measured limits rather than in isolation.

The lower accuracy of multiple-response items has been observed repeatedly across examination systems. Sadeq et al [19] found that tested models lagged consistently on choose-N-from-many items on a UK medical examination. Tang et al [20] reported a significant question-type main effect but no language main effect on a Chinese licensing examination. Zong et al [7], spanning 5 years and the physician, pharmacist, and nurse populations, showed the same pattern. Each, however, stopped at overall accuracy, neither splitting errors into omission and commission nor examining the source of the gap.

A few studies addressed classification of multiple-response errors but did not separate error direction at the option level. Gérard et al [9] treated omissions and false selections as equally weighted deductions under partial credit. Du et al [10] merged under- and overselection into a single incorrect category. Goodings et al [21], although classifying errors, obtained only reasoning-level labels such as logical and informational fallacies, which likewise cannot characterize the direction of option-level errors.

Our methodological advance is to standardize error direction by opportunity. With roughly 2.3 correct options for every incorrect option on X-type items, omission counts naturally dominate even under unbiased per-option judgment. Prior studies recorded both error types but did not control for this opportunity imbalance and therefore could not separate the structural artifact from genuine behavioral tendency. Using 2 independent routes, an option-level GEE and the OBI, we show consistently that the raw omission dominance is a product of the answer-key opportunity imbalance and that, once the opportunity base is controlled, the direction reverses toward commission.

These findings refine the bias spectrum reported by SATA-Bench on general select-all-that-apply tasks [13]. Across 32 proprietary and open-source models evaluated on a 1470-item human-validated set (more than 10,000 items released in total), that benchmark observed 3 coexisting biases. Models systematically avoided some correct options and included incorrect options under uncertainty, with nearly every model exceeding a 20% question-level false-positive rate. Stronger models speculated more, their false-positive rate correlating with exact-match accuracy at r=0.61, and they systematically underpredicted the number of correct options. SATA-Bench noted that under- and overselection can coexist, but its false-positive rate only records whether a question contains any incorrect option. It does not place omissions and commissions on the separate opportunity bases of correct and incorrect options and therefore cannot determine whether set-level underselection and per-opportunity false selection arise from answer-key structure or model behavior. We further show that the 2 can coexist within the same set of responses, where set-level underselection corresponds to answer-key structure and per-opportunity false selection is the behavioral tendency.

Studies of models aligned through posttraining procedures report a tendency to select rather than withhold when tested on counterfactual single-select items containing no valid option [22,23], a direction concordant with our higher per-opportunity commission rate. Those items contain no correct answer, whereas every X-type item does, so this remains suggestive context rather than evidence of an alignment mechanism. Risk-averse frameworks for medical AI treat acting with false certainty under uncertainty as the failure mode that caution modules and escalation to human review seek to suppress [24]. Our per-opportunity results show a commission-dominant error profile, which cautions against reading raw omission rates as evidence of excessive conservatism. Whether this profile reflects confidence miscalibration or implies a tuning direction cannot be determined from examination responses alone.

The item-type penalty and error direction were consistent across the 5 models and 2 languages. The item type × model interaction was nonsignificant at P=.19, and all pairwise between-model comparisons of strict omission were P≥.46 after Holm correction, in line with an independent study reporting common response patterns on medical MCQs across proprietary, open-source, and domain-specific models irrespective of architecture [25], indicating that the penalty and commission direction are not idiosyncratic defects of a single vendor or architecture.

The item type × language interaction was likewise nonsignificant, and its direction matches the absence of a Chinese to English difference on Chinese-origin items reported by Wu et al [26] for Chinese and US nursing licensure examinations. At the ensemble level, in an exploratory analysis, option-level majority voting raised X-type accuracy from 87% for the best single model to 89%, a margin not distinguishable from zero when paired on the item, consistent in direction with established gains from ensemble methods in medical question answering [1,14], and the 96% oracle bound shows that model errors are far from overlapping, leaving room for finer aggregation such as confidence weighting.

Implications

These results carry 2 direct implications for medical LLM evaluation. First, accuracy must be reported stratified by item type. Within this item bank, observed pooled accuracy was 12.52 percentage points lower on multiple-response than on single-select items (from 96.75% to 84.23%; 13.55 percentage points after adjustment for model, language, discipline, and cohort), and a pooled score dominated by single-select items systematically overstates performance on the more demanding multiple-response format. A meta-analysis of 36 studies was based almost entirely on single-best-answer items and likewise did not stratify its subgroup analysis by single-select vs multiple-response item type [27], which together with studies reporting only merged error rates [9,10] indicates that the reporting gap is not isolated. Second, multiple-response evaluation must report both the numerator and the denominator of each error type, or report opportunity-standardized error direction through an option-level model, since reporting only overall accuracy mistakes the answer-key opportunity imbalance for a behavioral feature of the model.

Strengths and Limitations

This study has several strengths. It drew on a large, authentic bank of 1644 items spanning 10 consecutive cohorts of a national medical examination, rather than newly assembled practice questions, and evaluated 5 mainstream LLMs bilingually under a standardized zero-shot protocol with web retrieval disabled, providing 16,440 standardized responses that allowed controlled within-item comparison. Methodologically, error direction was examined through 2 independent, mutually corroborating routes, namely, an option-level per-opportunity model and the OBI, and inference used cluster-robust GEEs supported by a prespecified program of sensitivity analyses, across which the principal findings remained stable.

Several limitations should nevertheless be noted. First, because the items came from a single national examination system, generalizability to other examinations, item formats, and scoring rules awaits replication. Second, the models reflect fixed versions at the time of testing, so absolute scores will shift with future updates. Third, only text-based zero-shot responses were evaluated; multimodal, interactive, and real-world clinical settings remain for future work. Fourth, because our items are publicly released and may appear in model training corpora, absolute accuracy levels may be inflated. Within-bank contrasts between item types, option types, and languages may also be affected if prior exposure or memorization differs across those comparisons. The cohort-recency diagnostic detected no statistically significant association but cannot exclude differential memorization or establish robustness to contamination. Fifth, option records within an item are not behaviorally independent, and clustering accommodates that dependence in variance estimation only; the answer-key-size stratification and interaction analysis assess sensitivity to key size but do not identify the structure of within-item dependence or model the joint selection process.

Conclusions

LLMs suffer a large performance penalty on medical multiple-response items that is consistent across models and languages. The omission dominance suggested by raw counts is an artifact of the answer-key opportunity imbalance, that is, the larger number of correct than incorrect options, and after opportunity standardization the direction of model errors reverses toward commission. Future medical LLM benchmarks should report accuracy stratified by item type and error direction in a manner that accounts for opportunity imbalance, on both the absolute and the relative scale. In this benchmark, option-level majority voting did not outperform the best single model by a margin distinguishable from zero, and whether option-level aggregation can reliably exceed the best single model remains open.

Acknowledgments

The authors thank Professor Wang for translating the Chinese examination items into English and Chief Physician Cui for independently checking the accuracy of the translations.

Generative AI Use

Generative AI tools were used solely for English-language polishing of this manuscript. All AI-assisted revisions were manually reviewed by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

Funding

The authors received no specific funding for this work.

Data Availability

The examination items are copyrighted works of China's National Postgraduate Entrance Examination and were used in this study for scientific research under the principle of fair use. The authors, as individuals, have no right to redeposit or redistribute the full item bank. All raw model responses are included in Multimedia Appendix 7. The analysis code is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: XR, HY
Data curation: YY, SD

Formal analysis: XR, XL
Methodology: HY, YL
Project administration: YL
Resources: YY, ML
Supervision: YL
Validation: SD, ML
Visualization: ML
Writing – original draft: XR, HY
Writing – review & editing: all authors
XR and HY are co-first authors and contributed equally. ML and YL are co-corresponding authors. All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study design, sample flow, item classification, and model evaluation protocol.
[DOCX File (Microsoft Word File), 877 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Chinese-English translation examples, 1 complete example per item type.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Full verbatim Chinese and English prompts used for all models and all items.
[DOCX File (Microsoft Word File), 14 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Prespecified sensitivity analyses and regression model diagnostics.
[DOCX File (Microsoft Word File), 783 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Response status, exclusions, and performance by analytic item type, examination cohort, and discipline.
[DOCX File (Microsoft Word File), 1155 KB-Multimedia Appendix 5]

Multimedia Appendix 6

X-type error patterns, option-level error rates, language comparisons, intermodel agreement, and majority-voting strategies.
[DOCX File (Microsoft Word File), 1728 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Item-by-item archive of all raw model responses.
[XLSX File (Microsoft Excel File), 1773 KB-Multimedia Appendix 7]

References

1. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
2. Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. Nov 6, 2024;10:e63430. [doi: 10.2196/63430] [Medline: 39504445]
3. Diao Y, Wu M, Xu J, Pan Y. Multiple large language models' performance on the Chinese medical licensing examination: quantitative comparative study. JMIR Hum Factors. Dec 16, 2025;12:e77978. [doi: 10.2196/77978] [Medline: 41401211]
4. Wang X, Long Z, Zhu B, et al. Evaluation of DeepSeek-R1 and ChatGPT-4o on the Chinese national medical licensing examination: a multi-year comparative study. Sci Rep. 2026;16(1):2237. [doi: 10.1038/s41598-025-31874-6]
5. Li KC, Bu ZJ, Shahjalal M, et al. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. PLoS ONE. 2024;19(4):e0301702. [doi: 10.1371/journal.pone.0301702] [Medline: 38573944]
6. Yu P, Fang C, Liu X, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ. Feb 9, 2024;10:e48514. [doi: 10.2196/48514] [Medline: 38335017]

7. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. Feb 14, 2024;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
8. Ming S, Guo Q, Cheng W, Lei B. Influence of model evolution and system roles on ChatGPT's performance in Chinese medical licensing exams: comparative study. JMIR Med Educ. Aug 13, 2024;10:e52784. [doi: 10.2196/52784] [Medline: 39140269]
9. Gérard AO, Merino D, Labriffe M, et al. Evaluating and leveraging large language models in clinical pharmacology and therapeutics assessment: from exam takers to exam shapers. Br J Clin Pharmacol. Aug 2026;92(8):2574-2583. [doi: 10.1002/bcp.70137] [Medline: 40495266]
10. Du C, Pan Y, Ng C, et al. Three large language models demonstrate competitive performance in Traditional Chinese Medicine national medical licensing examinations over two years. Sci Rep. Apr 17, 2026;16(1):17942. [doi: 10.1038/s41598-026-49200-z] [Medline: 41998323]
11. Huang C, Sun Y, Liu W. A comparative study of the performance of different large language models in the Chinese National Pharmacist Licensing Examination. Front Med. 2026;13:1880914. [doi: 10.3389/fmed.2026.1880914] [Medline: 42519773]
12. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open. Mar 15, 2024;14(3):e080558. [doi: 10.1136/bmjopen-2023-080558] [Medline: 38490655]
13. Xu W, Cui S, Fang X, Xue C, Eckman S, Reddy CK. SATA-bench: select all that apply benchmark for multiple choice questions. Presented at: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26). Aug 9-13, 2026:Association for Computing Machinery. 10104-10113; Jeju Island, Republic of Korea. 2026. [doi: 10.1145/3770855.3817472]
14. Lucas MM, Yang J, Pomeroy JK, Yang CC. Reasoning with large language models for medical question answering. J Am Med Inform Assoc. Sep 1, 2024;31(9):1964-1975. [doi: 10.1093/jamia/ocae131] [Medline: 38960731]
15. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. Apr 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
16. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. May 2008;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
17. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Machine Intell. Mar 1998;20(3):226-239. [doi: 10.1109/34.667881]
18. Bevizova K, Falougy HE, Thurzo A, Harsanyi S. Is virtual reality enhancing dental anatomy education? A systematic review and meta-analysis. BMC Med Educ. Nov 29, 2024;24(1):1395. [doi: 10.1186/s12909-024-06233-0] [Medline: 39614238]
19. Sadeq MA, Ghorab RMF, Ashry MH, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. Sci Rep. Aug 14, 2024;14(1):18859. [doi: 10.1038/s41598-024-68996-2] [Medline: 39143077]
20. Tang Y, Chen J, Wang S. Performance benchmarking of LLMs on Chinese national medical licensing education: cross-lingual and question-type effects. PLoS ONE. 2026;21(4):e0346518. [doi: 10.1371/journal.pone.0346518]
21. Goodings AJ, Kajitani S, Chhor A, et al. Assessment of ChatGPT-4 in family medicine board examinations using advanced AI learning and analytical methods: observational study. JMIR Med Educ. Oct 8, 2024;10:e56128. [doi: 10.2196/56128] [Medline: 39378442]
22. Góral G, Wiśnios E, Sankowski P. Wait, that's not an option: LLMs robustness with incorrect multiple-choice options. In: Budzianowski P, editor. Presented at: 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025); Jul 27 to Aug 1, 2025:1495-1515; Vienna, Austria. 2025.URL: https://aclanthology.org/2025.acl-long.75/ [Accessed 2026-11-09]
23. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun. Jan 14, 2025;16(1):642. [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
24. Thurzo A, Thurzo V. Embedding fear in medical AI: a risk-averse framework for safety and ethics. AI. 2025;6(5):101. [doi: 10.3390/ai6050101]
25. GriotM, VanderdonckтJ, YukselD, Hemptinne C. Pattern recognition or medical knowledge? The problem with multiple-choice questions in medicine. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025; Vienna, Austria. 2025.[doi: 10.18653/v1/2025.acl-long.266]
26. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on nursing licensure examinations in the United States and China: cross-sectional study. JMIR Med Educ. Oct 3, 2024;10:e52746. [doi: 10.2196/52746] [Medline: 39363539]

27. Nouri H, Mahdavi A, Abedi A, Mohammadnia A, Hamedan M, Amanzadeh M. Performance of large language models in medical licensing examinations: a systematic review and meta-analysis. J Educ Eval Health Prof. 2025;22(36):36. [doi: 10.3352/jeehp.2025.22.36] [Medline: 41248547]

Abbreviations

AC1: Gwet AC1
GEE: generalized estimating equation
LLM: large language model
MCQ: multiple-choice question
OBI: omission-bias index
OR: odds ratio
SATA: Select All That Apply

Edited by Taiane de Azevedo Cardoso; peer-reviewed by Andrej Thurzo, Varun Gazala; submitted 03.Sep.2026; initial comments to author 25.Sep.2026; final revised version received 20.Oct.2026; accepted 27.Oct.2026; published 23.Nov.2026

Please cite as:
Ran X, Yin H, Yang Y, Duan S, Luo X, Liu Y, Liao M
Error Direction of Large Language Models on Medical Multiple-Response Questions: Opportunity-Standardized Benchmark Study
JMIR Med Educ 2026;12:e111146
URL: https://mededu.jmir.org/2026/1/e111146
doi: 10.2196/111146

© Xinyan Ran, Haolin Yin, Yueming Yang, Shanshan Duan, Xia Luo, Yang Liu, Mengqian Liao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.