Accuracy Is Not Enough: Evaluating Response Stability of Small Open-Weight LLMs in Accounting Education

Susanto*[1], Jarot Budiasto[2], Okto Irianto[3]
[1]Department of Informatics Engineering, Universitas Musamus, Merauke, Indonesia
[2]Department of Information Systems, Universitas Musamus, Merauke, Indonesia
[3]Department of Accounting, Universitas Musamus, Merauke, Indonesia

E-mail: susanto@unmus.ac.id[1], jarot@unmus.ac.id[2], okto@unmus.ac.id[3]

Abstrak. Kecerdasan artifisial generatif semakin digunakan dalam pendidikan tinggi, tetapi evaluasi model bahasa besar (LLM) pada bidang akuntansi umumnya menekankan akurasi satu respons. Penelitian ini mengevaluasi tiga LLM open-weight instruction-tuned berukuran kecil—Qwen3-4B-Instruct-2507, Phi-4-mini-instruct, dan SmolLM3-3B—menggunakan 92 butir MMLU-Redux 2.0 Professional Accounting yang dianotasi sebagai ok dan asesmen akuntansi universitas independen berisi 20 butir. Setiap butir memperoleh satu respons deterministik dan lima respons stokastik terkontrol, dengan pairwise response agreement (PRA) untuk mengukur keterulangan. Dari 2,460 generasi, akurasi deterministik utama masing-masing sebesar 56.52%, 54.35%, dan 40.22%. Akurasi Qwen dan Phi tidak berbeda signifikan, tetapi PRA Qwen jauh lebih tinggi (0.946 dibanding 0.768), sedangkan SmolLM3 mencapai 0.655; seluruh perbedaan PRA berpasangan tetap signifikan setelah koreksi Holm. Stabilitas tinggi tidak menjamin kebenaran: Qwen salah pada 40.22% butir utama dalam seluruh lima pengulangan stokastik. Asesmen eksternal kembali menunjukkan stabilitas Qwen yang lebih tinggi dan akurasi numerik yang secara deskriptif lebih rendah daripada naratif untuk seluruh model, tetapi perbedaan subkelompok tidak signifikan setelah koreksi multipel. Hasil ini menunjukkan bahwa evaluasi LLM untuk pendidikan akuntansi perlu mengukur keterulangan respons bersama dengan kebenaran jawaban.

Kata kunci: pendidikan akuntansi; model bahasa besar; stabilitas respons; reliabilitas; asesmen pilihan ganda

Abstract. Generative artificial intelligence is increasingly used in higher education, but accounting evaluations of large language models (LLMs) usually emphasize single-response accuracy. This study evaluates three small open-weight instruction-tuned LLMs—Qwen3-4B-Instruct-2507, Phi-4-mini-instruct, and SmolLM3-3B—using 92 MMLU-Redux 2.0 Professional Accounting items annotated as ok and an independent 20-item university accounting assessment. Each item received one deterministic response and five controlled stochastic responses, with pairwise response agreement (PRA) measuring repeatability. Across 2,460 generations, primary deterministic accuracy was 56.52%, 54.35%, and 40.22%, respectively. Qwen and Phi did not differ significantly in accuracy, but Qwen showed substantially higher PRA (0.946 vs. 0.768), while SmolLM3 reached 0.655; all pairwise PRA differences remained significant after Holm correction. High stability did not guarantee correctness: Qwen was wrong on 40.22% of primary items in all five stochastic runs. The external assessment again showed higher Qwen stability and descriptively lower numerical than narrative accuracy for all models, although subgroup contrasts were not significant after multiplicity correction. These results show that accounting-education evaluations should measure response repeatability alongside correctness.

Keywords: accounting education; large language models; response stability; reliability; multiple-choice assessment

1. Introduction

Generative artificial intelligence (GenAI) has moved rapidly from a general-purpose conversational technology into a practical component of higher education. In accounting, this transition is especially

consequential because students and professionals routinely work with tasks that combine conceptual knowledge, numerical procedures, rule application, and professional judgment. Early studies showed that ChatGPT could answer a substantial proportion of accounting assessment questions, but also documented uneven performance across topics and task types [1]-[4]. More recent work has emphasized that GenAI may support assessment design and foundational learning while remaining less dependable for higher-order tasks [5], and critical reviews have argued that accounting education should respond by strengthening critical engagement rather than treating AI output as authoritative [6], [7].

The educational question is therefore no longer only whether an LLM can produce a correct answer. Students may query the same system more than once, regenerate an answer, alter decoding conditions indirectly through an interface, or encounter stochastic output variation that is hidden behind an apparently deterministic conversational experience. A model that answers 55% of questions correctly but changes its answer frequently presents a different behavioral profile and verification challenge from a model with similar accuracy that repeats the same answer consistently. Conversely, repeatability does not imply correctness: a model may be confidently and consistently wrong. This distinction matters when students use AI for checking calculations, verifying an accounting treatment, preparing for examinations, or resolving uncertainty in self-study.

Most empirical accounting studies have assessed capability through one or a small number of responses. Wood et al. [1] evaluated 28,085 accounting questions and reported substantial but clearly imperfect ChatGPT performance. Abeysekera [2] compared numerical and narrative multiple-choice questions in introductory and advanced financial accounting. Eulerich et al. [3] showed that performance on professional accounting and auditing examinations improved markedly with stronger models, few-shot examples, and tools. Albuquerque and Santos [4] similarly found uneven results on a professional-access accounting examination, particularly for matters requiring judgment. These studies establish that accounting competence is model- and task-dependent, but they do not make run-to-run answer stability a central evaluation target.

Outside accounting, reliability-oriented research increasingly questions whether a single benchmark score is sufficient. MMLU-Redux was created after systematic re-annotation revealed nontrivial ground-truth and item-quality problems in the original MMLU benchmark [8]. Repeated-response studies have also shown that identical multiple-choice questions can yield variable LLM answers and that consistency can separate persistent correctness, volatility, and persistent error [9], [10]. Option-perturbation research further demonstrates sensitivity to answer position and formatting [11]. These findings motivate an accounting-specific evaluation in which correctness and response stability are measured separately under a controlled design.

This study addresses that gap using three small open-weight instruction-tuned LLMs and two accounting assessment sources. The primary evaluation uses MMLU-Redux 2.0 Professional Accounting, restricting inferential correctness analyses to the 92 items annotated as ok under the benchmark re-annotation scheme. An independent university-level dataset from Abeysekera [2] provides a small external assessment containing balanced introductory/advanced and numerical/narrative questions. Each item is evaluated once under deterministic decoding and five times under a predefined stochastic regime. Pairwise Response Agreement (PRA) is used as the principal measure of repeated-answer stability, complemented by all-run agreement, normalized entropy, correctness profiles, and format compliance. A fixed non-identity choice shuffle tests sensitivity to answer-option ordering without changing item content.

The study asks three research questions:

RQ1. How accurately do small open-weight LLMs answer professional and university-level accounting multiple-choice questions?

RQ2. How stable are small open-weight LLM responses under repeated controlled stochastic inference, and what information does response stability provide beyond accuracy alone?

RQ3. How does accounting-question performance vary across course level and question format in an independent university-level assessment set?

The study contributes in four ways. First, it evaluates small open-weight models using both correctness and repeated-response stability rather than a single benchmark score. Second, it provides direct evidence that similar accounting accuracy can coexist with materially different stability. Third, it

combines an annotation-filtered benchmark with explicit separation of problematic items and a fixed choice-order sensitivity analysis. Fourth, it examines introductory versus advanced and numerical versus narrative accounting questions on an independent assessment without pooling the two datasets. The central argument is deliberately narrow: accuracy and response repeatability capture different properties of LLM behavior, and both should be considered when evaluating learner-facing accounting applications.

2. Related Work

2.1 Generative AI in Accounting Education and Assessment

Accounting education was among the disciplines that rapidly tested ChatGPT against conventional assessments. Wood et al. [1] compared ChatGPT with students across 14 countries and 186 institutions and found that the model answered 56.5% of questions correctly, while students achieved higher assessment performance overall. Abeysekera [2] reported stronger performance in introductory than advanced financial accounting and examined numerical and narrative question forms. Evidence from professional examinations likewise shows that aggregate accuracy can conceal substantial variation across accounting tasks [3], [4].

Evidence from professional assessments reinforces that conclusion. Eulerich et al. [3] reported that baseline ChatGPT-3.5 did not pass the evaluated accounting certification exams, whereas GPT-4, few-shot examples, and reasoning-and-acting tools produced large performance gains. Albuquerque and Santos [4] found that ChatGPT came close to, but did not pass, a Portuguese professional-access examination and was less accurate in areas requiring professional judgment. Thus, model version and inference resources are integral to measured capability, making reproducible protocol specification important when interpreting benchmark scores.

The educational implications extend beyond exam integrity. Tharapos et al. [5] found stronger utility for lower- to mid-order learning tasks than higher-order contexts. Ballantine et al. [6] argued for accounting education that uses AI as an opportunity to re-emphasize critical and human dimensions of the discipline. Dosumu et al. [7] described how accounting assessment practice was already adapting to GenAI, while Sundkvist and Kulset [12] found that students recognized that ChatGPT answers could not always be trusted. At the broader higher-education level, a scoping review and systematic reviews highlight both potential benefits and risks of overreliance and cognitive offloading [13]-[15]. These concerns make it important to evaluate not only whether an answer is correct, but also how predictable model behavior is under repeated use.

2.2 Reliability Beyond Single-Response Accuracy

A conventional multiple-choice benchmark reduces each item-model interaction to correct or incorrect. This is efficient for capability ranking but discards information about output variability. Gema et al. [8] show that benchmark re-annotation can change model performance estimates. In medical education, Hanss et al. [10] evaluated ten responses per question and examined response consistency alongside accuracy, while Gürses and Ceylan [9] distinguished consistently correct, volatile, and consistently wrong behavior. These studies support treating repeatability as a separate behavioral dimension rather than as a synonym for correctness.

Stability should not be treated as a synonym for factual reliability. A model can repeat an incorrect answer perfectly, which yields high consistency but zero correctness for that item. For this reason, the present study operationalizes response stability as repeatability under a fixed stochastic inference condition, not as a claim that the output is trustworthy, safe, or pedagogically correct. Correctness and stability are reported side by side, and correctness profiles explicitly distinguish consistently correct from consistently wrong items.

2.3 Multiple-Choice Robustness and Answer-Order Sensitivity

Multiple-choice evaluations may also be sensitive to representation choices. Zhou et al. [11] showed that option-position shuffling and other knowledge-equivalent transformations can change model answers, particularly for smaller models. This creates a potential confound when option positions

interact with learned response preferences. The present study therefore includes a fixed answer-choice shuffle as a sensitivity analysis. The purpose is not to diagnose training-data contamination; an answer change can reflect position bias, decoding behavior, benchmark familiarity, or other factors. The sensitivity test asks only whether aggregate accuracy and selected semantic options are robust to a fixed non-identity permutation.

2.4 Research Gap

The accounting literature has established that LLM performance varies by model, task, assessment level, and access to tools, while the broader LLM-evaluation literature increasingly treats repeatability and benchmark quality as distinct concerns. A recent scoping review also documents rapid growth in accounting and finance research on ChatGPT across educational, professional, and research applications [16]. What remains less developed is a compact accounting-education evaluation that combines annotation-filtered benchmark scoring, repeated stochastic inference, item-level stability statistics, a representation-sensitivity check, and an independent university-level assessment set. The contribution is therefore an accounting-specific extension of stability-aware LLM evaluation rather than a new general stability metric or a claim of general-purpose model superiority.

3. Methodology

3.1 Study Design

The study used a two-part design. The primary evaluation measured accounting accuracy and response stability on MMLU-Redux 2.0 Professional Accounting, while the external assessment used the 20 multiple-choice questions released with Abeysekera [2]. The two datasets were analyzed separately; no pooled accuracy was calculated. Three small open-weight instruction-tuned models were tested under the same zero-shot prompt and decoding settings. The study produced 2,460 responses: 2,100 from the primary evaluation, including the fixed choice-order sensitivity condition, and 360 from the external assessment. Table 1 summarizes the design.

Table 1. Experimental design and datasets

Component	Dataset	Items	Role	Inference per item/model
Primary	MMLU-Redux 2.0 Professional Accounting	100 total; 92 annotated ok	Professional accounting MCQs	1 deterministic + 5 stochastic; plus fixed shuffle deterministic
External	Abeysekera (2024)	20	10 introductory / 10 advanced; 10 numerical / 10 narrative	1 deterministic + 5 stochastic

3.2 Primary Dataset and Quality Control

MMLU-Redux was developed by manually re-annotating MMLU questions and assigning item-level error categories [8]. The Professional Accounting configuration contains 100 items. All 100 were used for inference, but primary correctness analyses were restricted to the 92 items whose error_type was ok. The remaining eight items carried annotations such as no correct answer, expert-review flags, or option-clarity problems and were excluded from primary accuracy estimation. Thus, the primary results are based on the 92 items without an identified annotation issue under the MMLU-Redux re-annotation scheme.

The analysis used MMLU-Redux 2.0 dataset commit 372ea425445d51e1ba1188c56e5e893f8138621f. Item-level content hashes were recorded to verify dataset identity across runs.

3.3 External University-Level Assessment

The external dataset comprised the 20 questions accompanying Abeysekera [2]: 10 Introductory Financial Accounting items and 10 Advanced Financial Accounting items, with five numerical and five

narrative questions at each level. The public research-data PDF was converted to a machine-readable dataset. Structural checks confirmed 20 unique items, four non-empty options per item, balanced course-level and question-format counts, and one A-D source answer label per item. Fidelity checks verified the conversion against the published source; they did not independently re-adjudicate the substantive accounting correctness of all 20 published answer keys.

One source inconsistency was retained explicitly. For Question 6, the source summary table lists D, whereas the item body labels B as the correct answer. The item-body label B was used in the main analysis. Sensitivity analyses also evaluated exclusion of Question 6 and scoring it as D.

3.4 Models and Reproducible Runtime

The model set was limited to instruction-tuned open-weight models small enough to run sequentially in FP16 on a single Kaggle Tesla T4. Qwen3 represents a recent dense model family with strong general reasoning and instruction-following capabilities [17]. Phi-4-mini is a 3.8-billion-parameter model designed for compact reasoning workloads [18]. SmolLM3-3B provides a third 3-billion-parameter comparison point and supports reasoning and non-reasoning instruction modes [19]. Exact repository commit identifiers were recorded for all three models; none was fine-tuned or quantized for this study.

All evaluations used Python 3.12, PyTorch 2.10.0+cu128, Transformers 4.57.6, CUDA 12.8, and one Tesla T4 GPU. Models were loaded one at a time in FP16. Phi-4-mini-instruct used the native Transformers Phi3ForCausalLM implementation, and SmolLM3 thinking mode was disabled.

3.5 Prompt and Inference Protocol

Each question was presented in zero-shot form with four labeled options. The system instruction was: "You are completing an accounting multiple-choice assessment. Select the single best answer. Respond only with one uppercase letter: A, B, C, or D. Do not provide an explanation." No chain-of-thought prompting, retrieval-augmented generation, few-shot demonstrations, calculators, external tools, or self-verification were used. This isolates direct-answer behavior and keeps inference comparable across models.

The deterministic condition used greedy decoding (do_sample=False, num_beams=1, max_new_tokens=8). The stochastic condition used do_sample=True, temperature 0.7, top-p 0.90, top-k 0, typical-p 1.0, repetition penalty 1.0, one beam, and eight maximum new tokens. Each item received five stochastic generations with base seeds 101, 202, 303, 404, and 505. Seeds were combined with the item index using a fixed formula. Native end-of-sequence configuration was preserved. A conservative answer parser accepted exact A-D outputs and explicitly delimited leading letters, including a single answer letter followed by a newline and explanatory text. Strict one-letter format compliance was recorded separately from lenient answer validity.

3.6 Response-Stability Metrics

For item i, let r_{ia} denote its response in stochastic run a across R repeated runs. Pairwise Response Agreement (PRA) was defined formally as:

$$\mathrm{PRA}_i = \frac{1}{\binom{R}{2}} \sum_{1 \le a < b \le R} \mathbb{I}\,(r_{ia} = r_{ib}), \quad R = 5 \tag{1}$$

where $\mathbb{I}(\cdot)$ equals 1 when two responses are identical and 0 otherwise. With $R = 5$, $C(5,2) = 10$ unordered response pairs are compared per item: AAAAA gives PRA=1.00, AAAAB gives PRA=0.60, and AAABB gives PRA=0.40. Higher PRA means greater run-to-run repeatability, not greater correctness. Invalid outputs, if present, were retained as a response state rather than dropped; all observed stochastic responses were parseable.

Secondary measures were all-run agreement (all five responses identical), normalized answer entropy over A-D/INVALID, five-run stochastic accuracy, and the item-level number of correct responses (0/5 through 5/5). Together, these measures distinguish persistent correctness, response volatility, and persistent error.

3.7 Statistical Analysis
All inferential tests were two-sided with alpha=.05. The assessment item was the statistical unit; the five stochastic generations were repeated observations within an item, not independent samples. For RQ1, deterministic item-level correctness (0/1) was summarized by accuracy and Wilson 95% confidence intervals [20]. Because the same items were answered by all models, Cochran's Q tested the paired three-model omnibus difference [21]. The three planned pairwise comparisons used exact McNemar tests [22] with Holm correction [23], while 10,000 paired item-level bootstrap samples provided confidence intervals for accuracy differences [24].

For RQ2, each item contributed one PRA value per model. Friedman's rank test assessed the paired three-model omnibus difference [25]. The three planned pairwise mean-PRA differences used 10,000 two-sided sign-flip permutation tests with Holm correction and 10,000 paired-item bootstrap confidence intervals. For exploratory RQ3, deterministic introductory-versus-advanced and numerical-versus-narrative contrasts were tested within each model using two-sided Fisher exact tests because subgroup counts were small. The six tests (two contrasts x three models) formed a separate Holm-corrected family; the four crossed cells (n=5 each) were descriptive only.

Choice-order sensitivity paired original and shuffled deterministic correctness within item and model; exact two-sided McNemar tests were Holm-corrected across the three models, with paired bootstrap confidence intervals for accuracy changes. Bootstrap and permutation analyses used seed 2026. Paired-bootstrap 95% intervals are unadjusted effect-size intervals, whereas p-values are multiplicity-adjusted where stated.

3.8 Fixed Answer-Choice-Order Sensitivity Analysis
Each of the 100 MMLU Professional Accounting items received one deterministic non-identity permutation of the four options. A fixed SHA-256-derived permutation was used for each item and applied identically across models; the ground-truth label was remapped accordingly. Shuffled results were not pooled with original-order results. On the 92 primary items annotated as ok, original-versus-shuffled correctness was compared within each model using exact McNemar tests with Holm correction and paired bootstrap confidence intervals. A semantic option-flip rate compared the original selected option with the shuffled selection mapped back to its original option identity.

3.9 Research Ethics and Reproducibility
The study did not recruit students or other human participants and did not use personal or institutional student data. It evaluated publicly available assessment material and model outputs. Dataset and model commit identifiers, software versions, generation parameters, seeds, and raw outputs were recorded for reproducibility. Fixed commits and seeds improve operational reproducibility, although GPU generation is not guaranteed to be bitwise identical across different hardware and software stacks.

4. Results

4.1 Primary Deterministic Accuracy (RQ1)
On the 92 MMLU-Redux items annotated as ok, Qwen3-4B-Instruct-2507 obtained the highest deterministic point estimate (56.52%), followed by Phi-4-mini-instruct (54.35%) and SmolLM3-3B (40.22%). The omnibus Cochran test was significant (Q=9.476, p=.0088), indicating that the paired accuracy profiles were not identical. Table 2 reports these deterministic estimates and their Wilson intervals.

Table 2. Primary deterministic accuracy on 92 MMLU-Redux items annotated as ok

Model	Correct / N	Accuracy	Wilson 95% CI
Qwen3-4B-Instruct-2507	52/92	56.52%	46.33-66.19%
Phi-4-mini-instruct	50/92	54.35%	44.20-64.15%
SmolLM3-3B	37/92	40.22%	30.78-50.43%

Pairwise analysis shows why the omnibus result should not be converted into a simple ranking claim. Qwen and Phi differed by only +2.17 percentage points (95% paired-bootstrap CI -8.70 to +13.04; Holm-adjusted p=.839). Qwen exceeded SmolLM3 by +16.30 points (95% CI +5.43 to +27.17; Holm-adjusted p=.0178). Phi exceeded SmolLM3 by +14.13 points, but the exact McNemar comparison did not remain significant after Holm correction (Holm-adjusted p=.0702). Thus, the evidence supports Qwen's higher point estimate, not a claim that it significantly outperformed both alternatives in accuracy.

4.2 Response Stability Under Repeated Stochastic Inference (RQ2)

Using the item-level PRA defined in Section 3.6, the stability results produced much clearer separation than deterministic accuracy. Qwen's mean PRA was 0.946, compared with 0.768 for Phi and 0.655 for SmolLM3. Friedman's test was highly significant (χ^2=51.616, p=6.19×10^{-12}). Every planned pairwise PRA comparison remained significant after Holm correction: Qwen-Phi Δ=+0.177 (95% CI 0.117-0.238, $p_{Holm} = .00030$), Qwen-SmolLM3 Δ=+0.290 (95% CI 0.221-0.361, $p_{Holm} = .00030$), and Phi-SmolLM3 Δ=+0.113 (95% CI 0.035-0.190, $p_{Holm} = .00740$). Table 3 summarizes the stochastic accuracy and response-stability metrics.

Table 3. Stochastic accuracy and response-stability metrics

Model	Stochastic accuracy	Mean PRA	All-run agreement	Norm. entropy	Strict format
Qwen3-4B-Instruct-2507	57.17%	0.946	90.22%	0.039	100.0%
Phi-4-mini-instruct	55.65%	0.768	57.61%	0.177	20.65%
SmolLM3-3B	46.74%	0.655	44.57%	0.271	17.83%

Despite low strict one-letter compliance for Phi and SmolLM3, all stochastic outputs from all three models remained parseable under the specified answer parser; therefore, no stochastic prediction was dropped from accuracy or PRA calculations.

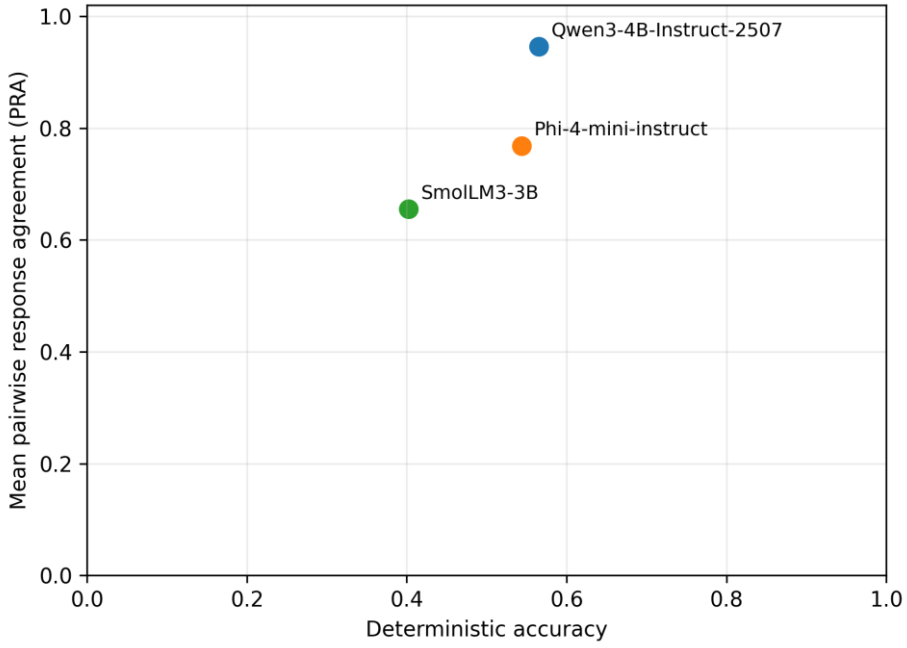


Figure 1. Deterministic accuracy versus mean pairwise response agreement on the primary annotation-filtered benchmark. Higher values on both axes are desirable, but the two dimensions are not interchangeable.

As shown in Figure 1, Qwen and Phi occupy similar positions on the accuracy axis but are separated substantially on the stability axis. A comparison that appears nearly tied from deterministic accuracy alone therefore changes when repeated-answer behavior is considered.

4.3 Stability Does Not Imply Correctness

High PRA should not be interpreted as evidence that repeated answers are correct. Qwen answered 49 of 92 items correctly in all five stochastic runs (53.26%), but it also answered 37 items incorrectly in all five runs (40.22%). Phi was correct in all five runs on 42.39% of items and wrong in all five on 31.52%. SmolLM3 was correct in all five on 28.26% and wrong in all five on 30.43%. Figure 2 shows the full 0/5-to-5/5 correctness profile for each model.

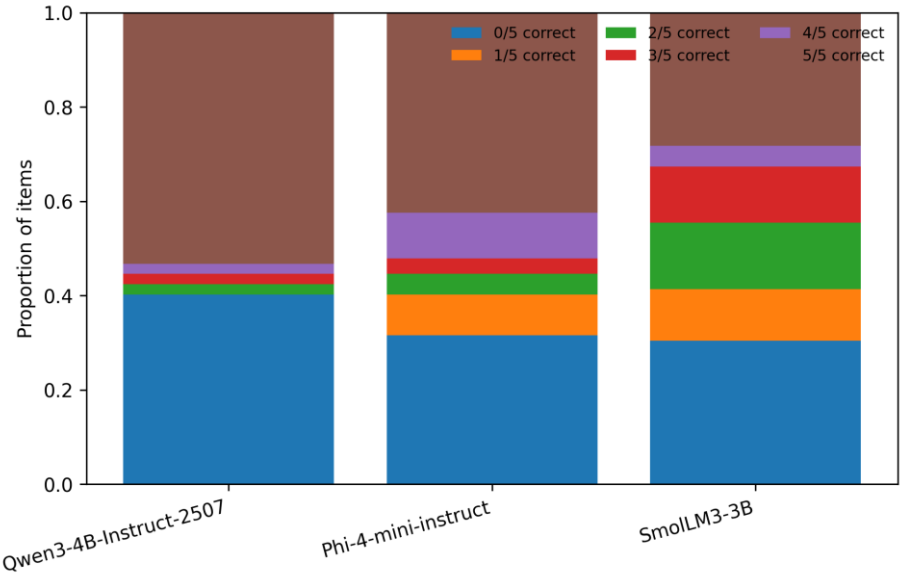


Figure 2. Distribution of item-level correctness across five stochastic runs. Large 0/5 and 5/5 segments show that stability can reflect either persistent error or persistent correctness.

The result is particularly important for educational interpretation. A student who regenerates an answer may encounter a model that is highly repeatable yet systematically wrong on a subset of questions. Stability is therefore a property of response behavior, not a substitute for correctness or verified accounting knowledge.

4.4 External University-Level Assessment (RQ1 and RQ3)

The 20-item external assessment produced deterministic accuracies of 55% for Qwen (11/20; Wilson 95% CI 34.21-74.18%), 50% for Phi (10/20; 29.93-70.07%), and 45% for SmolLM3 (9/20; 25.82-65.79%). Qwen again showed markedly higher response stability, with mean PRA=0.980 and 95% all-run agreement, compared with PRA=0.675 and 45% all-run agreement for Phi and PRA=0.670 and 40% all-run agreement for SmolLM3. Table 4 reports the overall and subgroup results. The small external accuracy differences should not be treated as robust model rankings.

Table 4. External assessment: overall and subgroup deterministic accuracy

Model	Overall	PRA	Intro.	Advanced	Numerical	Narrative
Qwen3-4B-Instruct-2507	55%	0.980	50%	60%	40%	70%
Phi-4-mini-instruct	50%	0.675	50%	50%	20%	80%
SmolLM3-3B	45%	0.670	60%	30%	30%	60%

Course-level differences were model-dependent: Qwen scored 50% on introductory and 60% on advanced items; Phi scored 50% on both; SmolLM3 scored 60% and 30%, respectively. No course-level Fisher comparison was significant after Holm correction. Therefore, the external data do not support a general claim that advanced items were uniformly more difficult.

Question-format results were directionally more consistent. All three models scored lower on numerical than narrative questions: Qwen 40% versus 70%, Phi 20% versus 80%, and SmolLM3 30%

versus 60%. Phi's raw Fisher p-value was .0230, but after the six-test Holm correction it became .1381; the other contrasts were also non-significant. The four course-level-by-format cells each contained only five items and were therefore treated descriptively. These results support only a tentative hypothesis that numerical accounting items may be more difficult under this direct-answer setting.

4.5 External Ground-Truth Sensitivity

Because Question 6 in the external source has conflicting published answer labels, overall external accuracy was checked under two alternatives. Using the item-body label B, Qwen/Phi/SmolLM3 accuracies were 55%/50%/45%. Excluding the item produced 52.63%/47.37%/47.37%, while using the summary-table label D produced 50%/45%/50%. Thus, one source ambiguity can materially affect ranking in a 20-item sample. The subgroup conclusion—no contrast surviving Holm correction—was unchanged.

4.6 Answer-Choice-Order Sensitivity

All three models showed lower point-estimate accuracy after the fixed option shuffle, but none of the changes was statistically reliable after Holm correction. Qwen decreased from 56.52% to 50.00% (Δ=-6.52 points; 95% CI -16.30 to +3.26; $p_{Holm} = .859$), Phi from 54.35% to 51.09% (Δ=-3.26; 95% CI -13.04 to +6.52; $p_{Holm} = .974$), and SmolLM3 from 40.22% to 34.78% (Δ=-5.43; 95% CI -17.39 to +6.52; $p_{Holm} = .974$). Table 5 summarizes the original/shuffled estimates and semantic option-flip rates.

Table 5. Fixed answer-choice-order sensitivity on the 92 primary items annotated as ok

Model	Original	Shuffled	Delta	Semantic flip	Holm p
Qwen3-4B-Instruct-2507	56.52%	50.00%	-6.52 pp	31.52%	.859
Phi-4-mini-instruct	54.35%	51.09%	-3.26 pp	34.78%	.974
SmolLM3-3B	40.22%	34.78%	-5.43 pp	56.52%	.974

Semantic option selections nevertheless changed on 31.52% of the 92 primary items for Qwen, 34.78% for Phi, and 56.52% for SmolLM3. These flip rates indicate representation sensitivity, particularly for SmolLM3, but they do not establish benchmark memorization or training-data contamination. Aggregate accuracy loss and semantic answer changes answer different questions and are therefore reported separately.

5. Discussion

5.1 Accuracy Does Not Fully Describe Model Reliability

The main finding is the divergence between correctness and response repeatability. Qwen and Phi were nearly tied on deterministic primary accuracy (56.52% versus 54.35%), and their pairwise difference was not significant. Yet Qwen's PRA exceeded Phi's by 0.177 and the difference remained highly significant after multiplicity correction. On the external assessment, deterministic accuracy differed by only five percentage points between Qwen and Phi, whereas Qwen again showed substantially higher PRA (0.980 versus 0.675). This cross-dataset contrast is the strongest evidence for the study's central claim that a single accuracy estimate can obscure meaningful differences in model behavior.

This result is consistent with reliability research beyond accounting. Repeated MCQ studies show that answer consistency varies by model and that apparent capability can conceal volatility [9], [10]. The present study adds domain-specific evidence for accounting and shows that the difference is visible even among relatively small models under an identical prompt, hardware environment, and decoding specification. For learner-facing settings, this matters because repeated agreement may be interpreted as confirmation. Repeatability is therefore informative only when considered alongside correctness.

5.2 Stable Wrong Answers Are an Educationally Important Failure Mode

Qwen's high stability came with a large consistently-wrong subset: 40.22% of primary items were answered incorrectly in all five stochastic runs. This makes clear why PRA must not be labeled "factual reliability." A model can be repeatable because its preference for an incorrect option is strong. Gürses and Ceylan [9] similarly identify consistently wrong behavior as a distinct failure mode. In accounting education, verification should therefore emphasize independent reasoning, authoritative standards or worked solutions, and error analysis rather than repeated prompting alone.

This interpretation also connects with educational concerns about overreliance. Systematic reviews describe a tension between GenAI efficiency and the outsourcing of cognitive or metacognitive work [14], [15]. In accounting, where procedural accuracy and professional judgment matter, high answer consistency may increase trust even when the underlying answer is wrong. Educational tasks can counter this by requiring students to justify, reconcile, or evaluate AI-generated outputs.

5.3 Numerical Accounting Appears More Challenging, but the Evidence Is Exploratory

All three models scored lower on numerical than narrative questions in the external dataset. This direction aligns with prior accounting evidence that numerical and complex accounting tasks can challenge conversational models [2], and with findings that performance can improve when stronger models, examples, calculators, or other tools are permitted [3]. The present direct-answer design intentionally excluded such tools, and each external format group contained only ten questions. After Holm correction, none of the numerical-versus-narrative contrasts was significant, so the pattern should be treated as exploratory.

The level comparison was even less systematic: Qwen performed somewhat better on advanced items, Phi was unchanged, and SmolLM3 performed worse. Course level alone may be too coarse a difficulty proxy because advanced items can vary in computational burden, ambiguity, professional judgment, and familiarity. Future accounting benchmarks should annotate the cognitive and procedural demands of each item instead of relying only on course labels.

5.4 Benchmark Quality and Representation Sensitivity

The decision to use MMLU-Redux rather than unfiltered MMLU was consequential. Gema et al. [8] show that benchmark label and item errors can materially affect model estimates. In the Professional Accounting configuration, eight of 100 items carried an annotation other than ok, so primary correctness analysis used the remaining 92 items. The external source also contained one answer-key inconsistency that changed small-sample rankings under alternative scoring. Benchmark quality is therefore part of model evaluation rather than a minor preprocessing detail.

The choice-order sensitivity analysis adds a second form of caution. The fixed shuffle produced non-significant aggregate accuracy decreases, yet many semantic selections changed, especially for SmolLM3. This is consistent with prior evidence that multiple-choice outputs can depend on option representation [11]. Because one permutation cannot identify the mechanism, the result is intentionally narrow: selected answers were sensitive to a non-semantic option-order change, but the analysis does not diagnose memorization or contamination.

5.5 Implications for Accounting Education

Three practical implications follow. First, educators evaluating candidate AI tools should not rely on a one-pass accuracy demonstration; repeated inference can reveal whether a model is stable, volatile, or consistently wrong on local accounting questions. Second, students should be taught that repeated agreement is not independent evidence. Verification should involve accounting logic, source documents, standards, or manual calculations. Third, assessment design can use AI limitations constructively through numerical and multi-step tasks, error diagnosis, comparison of alternative treatments, and justification of assumptions.

These implications are aligned with calls in accounting education to preserve reflective and critical work as GenAI becomes more capable [5]-[7], [12]. They do not imply that any evaluated model is a "reliable accounting tutor." The study measured direct MCQ correctness and repeatability under a

specific protocol; it did not measure student learning, misconception formation, transfer, or professional decision quality.

5.6 Limitations and Future Work

Several limitations bound the conclusions. First, the primary set contains 92 annotation-filtered items from one Professional Accounting benchmark configuration, and the external assessment contains only 20 questions. The external confidence intervals are therefore wide, and one ambiguous key materially changes model ordering. Second, all questions are in English and are multiple choice; results may not generalize to open-ended journal entries, financial-statement preparation, spreadsheets, standards interpretation, or multilingual accounting education. Third, only three 3-4B-class instruction-tuned open-weight models were compared. Larger proprietary or reasoning-specialized models may show different accuracy-stability trade-offs.

Fourth, stability was measured under one controlled stochastic regime with five repetitions. Different temperatures, sampling parameters, or interfaces may change PRA. Fifth, the study used direct answers without calculators, retrieval, or chain-of-thought; these restrictions improve comparability but do not represent every realistic tutoring workflow. Sixth, public benchmarks can be familiar to model training pipelines, and the choice-shuffle analysis cannot rule contamination in or out. Finally, no students participated, so the study cannot determine how response stability affects trust, overreliance, learning outcomes, or professional judgment. Future work should use larger expert-adjudicated accounting item banks, multiple stochastic regimes, time-stamped private assessments, open-ended and spreadsheet tasks, and controlled student studies that measure verification behavior and trust calibration.

6. Conclusion

This study evaluated three small open-weight LLMs on annotation-filtered professional accounting questions and an independent university-level accounting assessment using both accuracy and repeated-response stability. The results show that these metrics describe different aspects of model behavior. Qwen and Phi achieved similar deterministic accuracy on the primary benchmark, yet Qwen was substantially more stable under repeated stochastic inference; Qwen's higher response stability was also observed on the external assessment. At the same time, high stability did not guarantee correctness, as a substantial subset of items was answered consistently incorrectly. All three models also showed descriptively lower accuracy on numerical than narrative external questions, although the small subgroup contrasts were not significant after multiplicity correction. A fixed answer-order shuffle changed many semantic selections but did not produce statistically reliable aggregate accuracy losses. These findings support a simple evaluation principle for accounting education: repeated-response behavior should be measured alongside conventional accuracy, while stability itself must not be mistaken for correctness. For learner-facing use, the practical objective is not merely to select the model with the highest single score, but to understand when a model is correct, when it is volatile, and when it is consistently wrong.

Data and Code Availability

The study used publicly available MMLU-Redux 2.0 Professional Accounting data and the public research-data questions associated with Abeysekera [2]. Reproducibility materials include model and dataset commit identifiers, generation parameters, code, raw outputs, item-level stability metrics, and statistical tables used for the reported analyses.

Generative AI Disclosure

Generative AI tools were used for language editing and organizational assistance. The authors verified and take responsibility for the study design, code, statistical analyses, interpretations, citations, and final text.

References

[1] D. A. Wood et al., “The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?” Issues in Accounting Education, 2023, doi: 10.2308/ISSUES-2023-013.

[2] I. Abeysekera, "ChatGPT and academia on accounting assessments," Journal of Open Innovation: Technology, Market, and Complexity, vol. 10, no. 1, Art. 100213, 2024, doi: 10.1016/j.joitmc.2024.100213.
[3] M. Eulerich, A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood, "Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries," Review of Accounting Studies, vol. 29, pp. 2318-2349, 2024, doi: 10.1007/s11142-024-09833-9.
[4] F. Albuquerque and P. Gomes dos Santos, "Can ChatGPT Be a Certified Accountant? Assessing the Responses of ChatGPT for the Professional Access Exam in Portugal," Administrative Sciences, vol. 14, no. 7, 2024, doi: 10.3390/admsci14070152.
[5] M. Tharapos et al., "Generative AI in Accounting Education: Evaluating ChatGPT's Role in Assessment and Skill Development," Accounting & Finance, 2025, doi: 10.1111/acfi.70051.
[6] J. Ballantine, G. Boyce, and G. Stoner, "A critical review of AI in accounting education: Threat and opportunity," Critical Perspectives on Accounting, 2024, doi: 10.1016/j.cpa.2024.102711.
[7] O. Dosumu, V.-A. Porumb, A. Stafford, and A. Zimmer, "In the wake of ChatGPT: early reflections on marking open-book online accounting assessments," Accounting Education, vol. 35, no. 4, pp. 425-456, 2026, doi: 10.1080/09639284.2025.2487487.
[8] A. P. Gema et al., "Are We Done with MMLU?" in Proc. NAACL 2025, pp. 5069-5096, 2025, doi: 10.18653/v1/2025.naacl-long.262.
[9] O. A. Gürses and I. Ceylan, "Consistency over accuracy: run-to-run stability of contemporary large language models on Turkish curriculum-aligned theoretical anatomy multiple-choice questions," BMC Medical Education, vol. 26, 2026, doi: 10.1186/s12909-026-08656-3.
[10] K. E. Hanss et al., "Assessing the Accuracy and Reliability of Large Language Models in Psychiatry Using Standardized Multiple-Choice Questions: Cross-Sectional Study," Journal of Medical Internet Research, vol. 27, 2025, doi: 10.2196/69910.
[11] W. Zhou, Q. Wang, M. Xu, M. Chen, and X. Duan, "Revisiting the Self-Consistency Challenges in Multi-Choice Question Formats for Large Language Model Evaluation," in Proc. LREC-COLING 2024, Torino, Italy, pp. 14103-14110, 2024.
[12] C. H. Sundkvist and E. M. Kulset, "Teaching accounting in the era of ChatGPT - The student perspective," Journal of Accounting Education, Art. 100932, 2024, doi: 10.1016/j.jaccedu.2024.100932.
[13] P. Newton and M. Xiromeriti, "ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review," Assessment & Evaluation in Higher Education, vol. 49, pp. 781-798, 2024, doi: 10.1080/02602938.2023.2299059.
[14] Y. Qian, "Pedagogical Applications of Generative AI in Higher Education: A Systematic Review of the Field," TechTrends, vol. 69, pp. 1105-1120, 2025, doi: 10.1007/s11528-025-01100-1.
[15] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review," Smart Learning Environments, vol. 11, 2024, doi: 10.1186/s40561-024-00316-7.
[16] M. M. Dong, T. C. Stratopoulos, and V. X. Wang, "A scoping review of ChatGPT research in accounting and finance," International Journal of Accounting Information Systems, vol. 55, Art. 100715, 2024, doi: 10.1016/j.accinf.2024.100715.
[17] A. Yang et al., "Qwen3 Technical Report," arXiv:2505.09388, 2025, doi: 10.48550/arXiv.2505.09388.
[18] A. Abouelenin et al., "Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs," arXiv:2503.01743, 2025, doi: 10.48550/arXiv.2503.01743.
[19] E. Bakouch et al., "SmolLM3: smol, multilingual, long-context reasoner," Hugging Face Blog, Jul. 8, 2025. [Online]. Available: https://huggingface.co/blog/smollm3.
[20] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," Journal of the American Statistical Association, vol. 22, no. 158, pp. 209-212, 1927.
[21] W. G. Cochran, "The comparison of percentages in matched samples," Biometrika, vol. 37, nos. 3-4, pp. 256-266, 1950.
[22] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, pp. 153-157, 1947.
[23] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65-70, 1979.
[24] B. Efron, "Bootstrap methods: another look at the jackknife," The Annals of Statistics, vol. 7, no. 1, pp. 1-26, 1979, doi: 10.1214/aos/1176344552.
[25] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675-701, 1937.