Methodological Research Article

Beyond Answer-Key Accuracy: Provenance-Aware Evaluation of LLM Multiple-Choice Reasoning

Cuauhtémoc G. Gómez-D[1]* and Margarita García Nieto[2]

[1] ARTIFEX Consulting; Doctoral Program in Technology and Innovation Management, Universidad Tecnológica Latinoamericana en Línea (UTEL), Mexico. ORCID: 0009-0000-9133-226X; CVU: 721181.

[2] CECAPMKG Consulting; Doctoral Program in Technology and Innovation Management, Universidad Tecnológica Latinoamericana en Línea (UTEL), Mexico. ORCID: 0009-0006-4335-5145; CVU: 2112701.

*Corresponding author: ggomez@artifex-consulting.io

Abstract

Multiple-choice question answering is used to compare large language models (LLMs), yet answer-key agreement alone can conceal weak provenance, ambiguous items, post-hoc adaptation, and unsupported claims about internal model mechanisms. This methodological study audits a single educational case concerning Lewin's change model. The source artifact reported answers from three named systems, a platform key, simulated outputs for additional systems, and model-specific equations and hallucination scores. We classified every claim by evidential provenance, adjudicated the item against primary and historiographic sources, and formalized a dual-reference evaluation that separates answer-key conformity from evidence compatibility. The reanalysis shows that *Moving* is defensible as the keyed stage, while the wording combines Lewin's change sequence with Kelman's later mechanisms of identification and internalization. The source supports only a documentary observation: two independently reported initial answers matched the key and one did not; a revised answer obtained after key disclosure is not an independent test. Simulated model comparisons, proprietary weights, hallucination scores, and convergence claims are not empirically identifiable from the available record. We propose a provenance-aware protocol integrating key alignment, expert adjudication, option-order stability, calibration, repeated runs, and mixed-effects analysis. The contribution is a claim-bounded framework for converting anecdotal LLM comparisons into reproducible educational evaluation.

Keywords: large language models, multiple-choice evaluation, benchmark validity, provenance, educational assessment, robustness

1. Introduction

Multiple-choice question answering (MCQA) has become a standard component of large language model evaluation because it is inexpensive, scalable, and apparently objective. Benchmarks such as MMLU operationalize broad knowledge and reasoning as selection among a finite set of alternatives [1]. This convenience, however, can create a false equivalence between selecting the keyed option and demonstrating stable, evidence-grounded reasoning. Contemporary evaluation research has shown that model rankings and item-level answers can change when option order, answer symbols, prompting procedures, or scoring rules are modified [2-6]. Explanations generated after an answer also cannot automatically be treated as faithful reports of the internal process that produced that answer [7]. The problem is particularly acute in educational settings. An examination key is an operational scoring standard, but it is not necessarily a complete epistemic standard. A keyed response may be pedagogically acceptable yet theoretically composite, outdated, or ambiguously worded. Conversely, a model may challenge an imperfect key for defensible reasons but still fail the operational task. Treating either outcome as sufficient obscures the construct being measured. Educational testing theory therefore requires an explicit interpretation of scores and evidence supporting their intended use, rather than assuming that a correct mark proves the target competence [8,9].

This paper reanalyzes an archival comparison of LLM responses to one item about the three-stage change model commonly associated with Kurt Lewin. The source document reports initial responses attributed to ChatGPT, a system labeled "Gemini 3.5 Pro," and DeepSeek-R1; a platform answer key; a revised DeepSeek-R1 response after

feedback; simulated responses for several additional systems; and mathematical expressions presented as model-specific inference functions and hallucination measures. The archival document provides a useful case because it contains both a genuine evaluative question and multiple layers of unsupported inference.

The study does not attempt to rank LLMs from a single item. Instead, it asks three methodological questions:

RQ1: Which claims in the archival comparison are directly supported by inspectable evidence, which are derived, and which are simulated or unsupported?

RQ2: How should answer-key conformity be distinguished from compatibility with the best available theoretical evidence?

RQ3: What minimum design and mathematical framework are required for a reproducible confirmatory benchmark of LLM reasoning in educational MCQA?

The contribution is a dual-reference, provenance-aware evaluation framework. It preserves the practical relevance of answer keys while preventing them from being treated as unquestionable truth, and it prevents generated rationales or speculative equations from being misrepresented as observations of proprietary model internals.

2. Theoretical Background

2.1 Multiple-choice evaluation and construct validity

An MCQ score is meaningful only to the extent that the item, response process, scoring rule, and intended interpretation align. The Standards for Educational and Psychological Testing treat validity as evidence supporting the interpretation and use of scores, not as an intrinsic property of an item or instrument [8]. Messick similarly framed validity as an integrated evaluative judgment involving content, substantive theory, response processes, structure, generalizability, external relations, and consequences [9]. For LLM evaluation, these principles imply that a key match is one indicator within a wider validity argument.

Large-scale LLM benchmarks often compress this argument into exact-match accuracy. MMLU was influential because it tested many subjects under a common format [1], but later work demonstrated that leaderboard positions can shift under minor benchmark perturbations [4]. Option order alone can change model responses [2,3], and selection behavior can depend on the symbol used for an option rather than its semantic content [5]. Incorrect or "none of the above" alternatives introduce a further metacognitive challenge: a robust system should detect when the offered set does not contain a defensible answer rather than merely select the least implausible option [6,10]. These findings motivate separate measurement of correctness, stability, and response-format robustness.

2.2 Answer-key conformity and epistemic compatibility

Operational scoring and epistemic adjudication answer different questions. Let the answer key represent the option that the educational platform awards credit. An expert reference standard instead represents the option or set of options that can be defended using the construct definition, primary sources, and accepted historiography. The two standards may agree, but agreement must be demonstrated rather than assumed.

The case item asks which stage is characterized by a change agent influencing collaborators to adopt new attitudes, values, and behaviors through identification and internalization. Lewin's 1947 work discussed change as a movement from one quasi-stationary equilibrium to another, including the need to "unfreeze" existing conditions and stabilize a new state [11]. The widely taught three-step label *unfreeze-change/refreeze* is a later simplification and institutionalization of a more nuanced field-theoretical account [12]. The central action described in the item - adoption of new conduct during transition - is compatible with the middle *changing* or *moving* stage.

The words *identification* and *internalization*, however, are not distinctive terminology from Lewin's three-step model. Kelman formalized compliance, identification, and internalization as processes of attitude change in 1958 [13]. The item therefore combines two conceptual traditions: Lewin's temporal stage of change and Kelman's mechanisms of social influence. This makes the wording theoretically composite, but it does not make the keyed option *Moving* indefensible. The appropriate scientific conclusion is not that the platform is the "supreme authority" or that the key is necessarily wrong; it is that the item requires dual documentation of its scoring key and its theoretical lineage.

2.3 Provenance and rationale faithfulness

A scientific comparison must identify the origin of each claim. A model's textual explanation is an observable output, but it is not direct access to hidden activations, training gradients, proprietary weights, or a faithful chain of causation. Turpin et al. showed that chain-of-thought explanations can omit decisive biases and therefore be systematically unfaithful [7]. Work on sycophancy similarly shows that models may alter answers to align with user-stated beliefs or feedback [14]. Consequently, post-hoc verbal explanations should be analyzed as responses, not as privileged evidence of internal computation.

Provenance also determines which statistics are admissible. A directly logged answer can support an item-level correctness indicator. A simulated answer assigned to a model that was never queried cannot support empirical accuracy. An equation invented by an analyst cannot identify a commercial model's decision function unless the parameters and mechanism are independently documented and estimable. A negative "hallucination score" has no scientific meaning without an operational definition, a measurement scale, data, and validation evidence.

3. Materials and Methods

3.1 Design

We conducted a documentary methodological audit of one source artifact dated July 19, 2026. The unit of analysis was not the LLM as a population-level system but each evaluative claim in the artifact: item wording, answer key, reported response, revised response, simulated response, model-mechanism statement, equation, numerical score, and conclusion. The design is a case study and evidence audit; it is not an experiment and does not support inferential claims about comparative model performance.

3.2 Source corpus and extraction

The corpus comprised: (a) one four-option educational item; (b) the platform statement identifying *Moving* as the correct answer; (c) initial answers and narrative explanations attributed in the source to three systems; (d) one answer revision after the platform key had been disclosed; (e) hypothetical outputs for other named systems; and (f) mathematical sections purporting to describe model-specific inference functions, hallucination errors, and convergence behavior.

No raw API responses, screenshots tied to verifiable timestamps, complete prompts, system instructions, model endpoint identifiers, software versions, sampling parameters, random seeds, token probabilities, or independent replication logs were included. The labels "GPT-4/GPT-4o" and "Gemini 3.5 Pro" are ambiguous as model identifiers; they are retained only as archival labels and not interpreted as validated endpoints.

3.3 Provenance coding

Each claim was assigned to one of four mutually exclusive provenance classes shown in Table 1. The classification follows a conservative rule: when the record could not distinguish a real execution from a narrative reconstruction, the claim was described as *reported in the source* rather than independently observed.

Table 1. Provenance classes used in the documentary audit.

Class	Operational definition	Admissible use	Example from the source
Reported	Explicitly present as an item, key, or attributed output in the artifact	Documentary description; item-level indicator with caveat	Platform key = Moving
Derived	Computed transparently from reported evidence without adding unobserved parameters	Descriptive reanalysis	Whether an initial reported answer matched the key
Simulated	Hypothetical behavior assigned to a system that was not documented as executed	Scenario discussion only; no performance claim	Expected answers for Claude, Llama, Mistral, Grok, or Copilot
Unsupported	Mechanistic or numerical claim without observable data, validated scale, or identifiable parameters	Exclude from results	Proprietary weights, hallucination values, stochastic differential equations

A claim was considered traceable when the artifact provided an explicit evidence pointer and the claim did not exceed what that evidence could establish. No inter-rater reliability statistic was calculated because the present reanalysis was prepared as a methodological reconstruction rather than a completed multi-coder study. A confirmatory study should use at least two independent coders, a codebook, adjudication rules, and agreement estimates such as Cohen's kappa [15].

3.4 Dual reference-standard adjudication

For item i, let y_i^{key} denote the platform key and let $\mathcal{A}_i^{expert}$ denote the set of answers judged defensible by an expert panel using the item stem, option set, construct map, primary literature, and historiographic analysis. The archival item was adjudicated against Lewin's 1947 account [11], the historical reconstruction by Cummings et al. [12], and Kelman's 1958 social-influence framework [13].

The adjudication rule was deliberately non-binary at the construct level. An answer could be compatible with the key, compatible with expert evidence, both, or neither. Item-quality concerns - such as conceptual blending - were recorded separately from response correctness. For this case, *Moving* was retained as a defensible answer, while the stem was flagged as a composite operationalization because it embeds Kelman's mechanisms within a Lewinian stage.

3.5 Mathematical framework

The source artifact conflated transformer attention with the probability of a generated token. These are related but distinct quantities. A generic decoder produces a token distribution from its hidden state:

$$P_\theta(w_t \mid w_{<t}, x) = \text{softmax}(W_o h_t)_{w_t}.$$

Scaled dot-product attention is instead:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

Neither equation identifies the proprietary decision weights of a particular hosted model. For an MCQ, the observed selection can be represented generically as:

$$\hat{y}_{mipr} = \underset{y \in \mathcal{Y}_i}{\text{argmax}} s_{\theta_m}(y \mid x_i, c_i, \pi_p, r),$$

where m indexes the model endpoint, i the item, p an option permutation or prompt variant, r a repeated run, x_i the item text, c_i the evaluation context, π_p the presentation transformation, and s_{θ_m} the model-dependent selection score. For black-box systems, s_{θ_m} is generally not directly identifiable; the scientifically accessible object is the output under a documented input and configuration.

Two binary indicators preserve the distinction between operational and epistemic standards:

$$K_{mipr} = \mathbb{1}\left[\hat{y}_{mipr} = y_i^{key}\right],$$
$$E_{mipr} = \mathbb{1}\left[\hat{y}_{mipr} \in \mathcal{A}_i^{expert}\right].$$

The dual-reference state is:

$$\mathbf{z}_{mipr} = \left(K_{mipr}, E_{mipr}\right) \in \{(0,0), (0,1), (1,0), (1,1)\}.$$

This representation prevents a key match from automatically being labeled epistemically correct and prevents a theoretically defensible dissent from being mislabeled as operational success.

For P option permutations, item-level response stability can be estimated as pairwise agreement:

$$S_{mi} = \frac{2}{P(P-1)} \sum_{p<q} \mathbb{1}\left[\hat{y}_{mip} = \hat{y}_{miq}\right].$$

When a calibrated probability $\hat{p}_{mi}$ for the keyed option is available, the Brier score is:

$$\text{BS}_m = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_{mi} - K_{mi})^2.$$

A confirmatory repeated benchmark can use a logistic mixed-effects model:

$$\text{logit}\Pr(K_{mipr} = 1) = \beta_0 + \beta_m + \gamma_p + u_i + v_r,$$

where u_i captures item heterogeneity and v_r captures run-level variation. Additional fixed effects can represent prompt form, option position, domain, difficulty, and whether "none of the above" is present. None of these aggregate quantities was estimated in the present single-item audit.

3.6 Analysis strategy

The analysis proceeded in five stages: (1) extract the inspectable claims; (2) assign provenance; (3) adjudicate the reference standard; (4) compute only logically available item-level indicators; and (5) identify the minimum evidence needed for any stronger claim. Figure 1 summarizes this workflow.

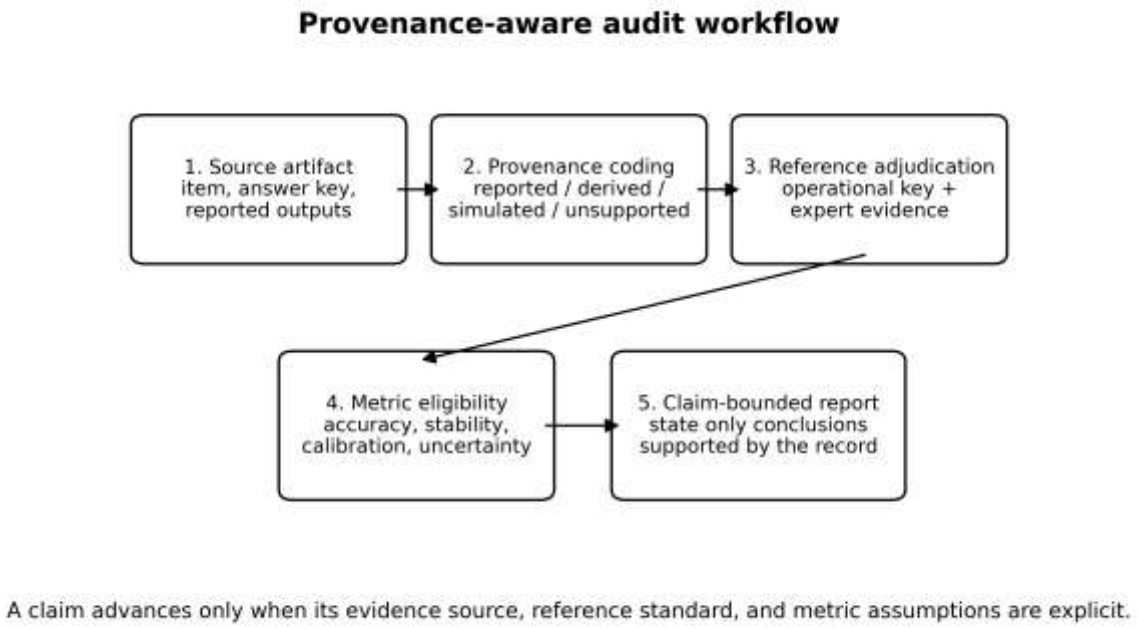

Figure 1. Provenance-aware audit workflow. A claim advances from the source artifact through provenance coding and reference-standard adjudication before a metric or conclusion is considered admissible.

4. Results

4.1 Claim-integrity audit

The source artifact contained a legitimate educational item and a platform answer key, but it mixed these with three qualitatively different forms of evidence. Table 2 reports the audit outcome. The most consequential problem was not a typographical equation error; it was a category error in which analytical inventions were presented as if they were measured properties of named systems.

Table 2. Scientific disposition of major claim types in the source artifact.

Claim type	Evidence available	Audit judgment	Required correction
Platform keyed Moving	Explicitly quoted in the artifact	Supported as an operational key	Retain; distinguish from epistemic truth
Initial answers attributed to three systems	Narrative report only; no raw execution logs	Usable as archival reported outputs	Label versions and provenance as unverified
Revised DeepSeek answer after feedback	Explicitly described as occurring after key disclosure	Supported as adaptation, not independent accuracy	Analyze separately from first-pass performance

Claim type	Evidence available	Audit judgment	Required correction
Simulated answers for additional systems	No documented executions	Not empirical	Remove from benchmark results
Model-specific weights and loss functions	No source code, telemetry, or parameter estimation	Unidentifiable	Replace with generic observable framework
Hallucination scores (e.g., 0.81 or -0.82)	No scale definition, ground truth, dataset, or validation	Scientifically uninterpretable	Remove
Stochastic dynamics and convergence theorem	No derivation tied to the systems or data	Unsupported formalism	Remove

The simulated section cannot be rescued by relabeling predictions as a "benchmark." It may motivate hypotheses, but it cannot estimate model performance. Likewise, arithmetic expressions that combine arbitrary numbers do not become measurements merely because they are written as equations. Measurement requires a defined construct, observable indicators, a scale, uncertainty, and validation evidence.

4.2 Reanalysis of the Lewin item

The item stem describes the adoption of new attitudes, values, and conduct under the influence of a change agent. This is temporally aligned with the middle stage of change rather than preparation (*unfreezing*) or stabilization (*refreezing*). The keyed answer *Moving* is therefore substantively defensible.

The source's contrary assertion - that "identification and internalization" necessarily make *None of the above* the theoretically correct option - is too strong. Those terms are associated with Kelman's processes of attitude change [13], not with Bandura or Bass as claimed in the source. Their inclusion indicates conceptual synthesis, but the stem still asks for a stage, and the behavior occurs during movement from the prior state to a new one. The scientifically appropriate item diagnosis is **key defensible; construct wording composite; citation lineage incomplete**.

This distinction matters because two errors would otherwise be possible. First, blindly accepting the key would ignore the composite wording. Second, rejecting *Moving* solely because two mechanisms originate outside Lewin would confuse mechanism attribution with stage classification.

4.3 Reported response matrix

Table 3 separates the first reported response from the answer produced after disclosure of the key. The labels are reproduced from the archival document and do not establish exact commercial endpoints.

Table 3. Documentary reanalysis of the reported responses to the single item.

Archival system label	Evaluation condition	Reported answer	Key alignment (K)	Evidence compatibility (E)	Interpretation
ChatGPT ("GPT-4/GPT-4o")	Initial, as reported	Moving	1	1	Key-aligned and defensible; endpoint ambiguous
"Gemini 3.5 Pro"	Initial, as reported	Moving	1	1	Key-aligned and defensible; endpoint unverified
DeepSeek-R1	Initial, as reported	None of the above	0	0	Rejects a defensible key on an overstated attribution claim
DeepSeek-R1	After platform feedback	Moving	1	1	Feedback-conditioned adaptation; not an independent trial

The only descriptive statement supported by the archival record is that two of the three *reported initial* answers matched the key on this one item. This is not an accuracy estimate: the denominator is one item per system, the

systems were not independently verified, and the run conditions are unknown. The post-feedback response demonstrates responsiveness to disclosed feedback, not first-pass correctness or learning in the statistical sense.

4.4 Why no model ranking or hallucination rate is estimable

A model ranking requires repeated observations over a defined item universe, comparable execution conditions, and a declared scoring rule. None is present. A hallucination rate additionally requires a unit of analysis and an adjudicated ground truth for factual claims. The source instead assigned scalar values without a validated coding protocol. Negative values were interpreted as "verified," although no meaningful measurement scale or threshold had been established.

The source also inferred causal mechanisms from narrative explanations. Such explanations may be useful for error taxonomy, but they cannot identify latent computational weights. In the absence of model telemetry or controlled interventions, statements such as "authority bias = 0.4" or "sequential-completion bias = 0.2" are non-identifiable. The corrected equations in Section 3.5 intentionally describe only a generic observational interface.

5. Proposed Dual-Reference, Provenance-Aware Evaluation

5.1 Evaluation dimensions

The proposed framework, termed **Dual-Reference, Provenance-Aware Evaluation (DRPAE)**, treats MCQA performance as a multidimensional measurement problem. Table 4 defines the minimum dimensions.

Table 4. DRPAE dimensions and recommended evidence.

Dimension	Question answered	Metric or evidence	Failure prevented
Key alignment	Did the output receive operational credit?	Exact-match indicator K with confidence interval	Treating anecdotal answers as benchmark accuracy
Evidence compatibility	Is the answer defensible under expert theory review?	Binary indicator E; adjudication notes; acceptable-answer set	Treating the key as unquestionable truth
Provenance	Was the claim observed, derived, simulated, or unsupported?	Claim ledger and evidence pointer	Mixing hypothetical and empirical results
Robustness	Is the answer invariant to irrelevant presentation changes?	Stability across permutations and prompt variants	Option-order and symbol bias
Calibration	Does confidence correspond to correctness?	Brier score, expected calibration error, reliability plot	Confidently wrong selections
Explanation quality	Does the rationale cite relevant evidence without fabricating mechanisms?	Expert rubric; source accuracy; contradiction checks	Treating generated rationale as hidden reasoning
Reproducibility	Can another team rerun the exact evaluation?	Prompt, endpoint, date, settings, raw outputs, code	Unverifiable vendor or version claims

DRPAE does not imply that every study must expose proprietary internals. Black-box evaluation is scientifically legitimate when inputs, outputs, settings, and constraints are reported precisely. The framework instead prohibits mechanistic certainty that exceeds the black-box evidence.

5.2 Confirmatory benchmark protocol

A confirmatory study should begin with an item bank sampled across content domains and difficulty strata. Every item should have: a construct specification; a keyed answer; an expert-adjudicated acceptable set; a rationale with

primary citations; distractor diagnoses; and a flag for composite or disputed wording. Items should be piloted before model comparison. An expert panel should independently rate answer defensibility and item clarity, with disagreements adjudicated and agreement reported.

Each model must be registered by exact endpoint or release identifier, provider, access date, interface, system prompt, temperature, top-p, maximum output tokens, tool availability, and other relevant settings. Closed systems should be evaluated within a bounded time window because silent updates may affect reproducibility. Each item should be executed under randomized option permutations, multiple semantically equivalent prompt templates, and repeated runs where sampling is non-deterministic. Raw prompts and completions should be stored immutably.

The study should predefine whether the scored response is the first option symbol, the final textual answer, or a parser-extracted choice. Recent work shows that these procedures can disagree [4,5]. Explanations should be evaluated separately from selections. A model should not receive additional correctness credit merely for producing a persuasive rationale after an incorrect choice.

Figure 2 presents the confirmatory pipeline.

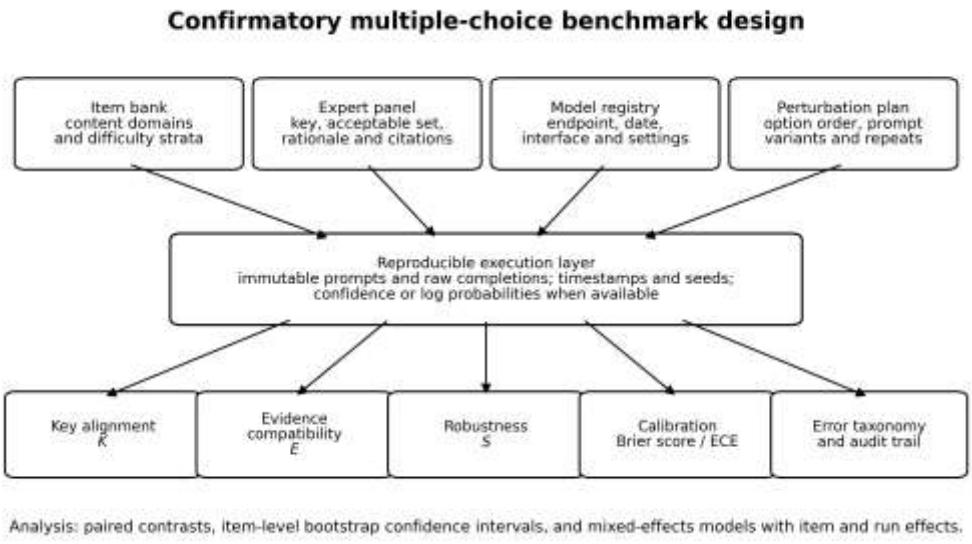

Figure 2. Confirmatory MCQ benchmark design integrating item construction, expert reference standards, model registration, perturbations, reproducible execution, and multidimensional analysis.

5.3 Statistical analysis and uncertainty

For a paired benchmark, model contrasts should be based on item-level paired outcomes rather than independent proportions. McNemar-type analyses can be used for two-model comparisons when assumptions are satisfied, while generalized linear mixed-effects models are preferable for multiple models, prompt forms, repeated runs, and clustered items. Bootstrap confidence intervals should be computed by resampling at the item level, preserving repeated observations within each item.

Sample size should be selected through simulation-based power analysis using plausible item heterogeneity, paired discordance, and the smallest practically important difference. A universal item count is not defensible

because power depends on these quantities and on the number of perturbations and repeats. Reporting should include effect estimates and uncertainty, not only p values or a ranked table.

Calibration analysis requires probabilities that correspond to the selected answer. When an API does not expose comparable probabilities, studies may collect explicit confidence judgments, but those judgments must be treated as model-generated reports and validated separately. Self-reported confidence is not equivalent to an internally normalized posterior.

5.4 Minimum reproducibility record

Table 5. Minimum record for an auditable LLM MCQ study.

Component	Required record
Item	Full stem, all options, original order, domain, source, license, construct map
Reference standard	Operational key, expert acceptable set, rationale, citations, adjudication date
Model	Provider, exact endpoint/version, access date, interface and region where relevant
Configuration	System/user prompts, temperature, top-, token limit, seed if supported, tools/RAG state
Execution	Run identifier, timestamp, option permutation, prompt variant, raw output, latency and errors
Scoring	Parser rules, missing/invalid response policy, key and evidence metrics, preregistered exclusions
Analysis	Code, package versions, confidence intervals, robustness and calibration procedures
Governance	Data permissions, privacy controls, conflict disclosure, AI-use statement

6. Discussion

6.1 Principal finding

The principal finding is methodological: the archival document is not a comparative LLM benchmark, although it contains the seed of a useful research question. It is a single-item documentary case that combines reported outputs, post-feedback adaptation, hypothetical simulations, and unsupported mathematical formalism. Scientific rigor is improved not by adding more equations, but by narrowing claims to the evidence and defining the conditions under which stronger claims would become testable.

The reanalysis also corrects the case's theoretical conclusion. *Moving* is a defensible answer to the Lewinian stage question. The presence of *identification* and *internalization* indicates that the item blends a Lewinian stage with Kelman's mechanisms of attitude change, but it does not logically force *None of the above*. The error lies in incomplete construct attribution and overly categorical adjudication, not necessarily in the platform key.

6.2 Implications for educational use

In educational practice, an LLM assistant faces at least three possible goals: predict the examiner's key, provide the best scholarly answer, or diagnose ambiguity. These goals should be made explicit. A high-quality response to a potentially composite item can state the likely keyed answer, explain why it fits the course framework, and note the theoretical caveat. This is superior both to uncritical compliance and to performative disagreement.

The framework therefore reframes "anti-sycophancy." Disagreement is not intrinsically rigorous, and agreement is not intrinsically sycophantic. The relevant property is whether the system calibrates its conclusion to evidence, identifies uncertainty, and resists unsupported user premises. In the case analyzed here, the initial DeepSeek answer

was not made more scientific merely by contradicting the key; its rationale relied on incorrect attribution and an overstrong inference.

6.3 Implications for LLM benchmarking

The case illustrates why one-shot demonstrations are vulnerable to narrative overfitting. Once an analyst sees a preferred answer, it is easy to assign each model a plausible personality - contextual, literal, constitutional, anti-hallucinatory - and then write an equation that encodes the story. Such equations are explanatory metaphors, not estimated models. The appropriate remedy is preregistration, raw logging, perturbation testing, and explicit separation of observed and simulated evidence.

Option-order robustness is especially important because MCQA ostensibly measures knowledge but can partially measure positional preference [2-5]. The use of "None of the above" adds another layer: systems may need to reject the offered set rather than select the least incorrect option [6,10]. A robust benchmark should therefore report both ordinary items and controlled adversarial variants.

6.4 Limitations

7. Conclusion

A scientifically defensible LLM evaluation must distinguish what was observed from what was inferred, simulated, or invented. In the analyzed educational case, the available evidence supports a narrow conclusion: two reported first-pass outputs selected the platform's defensible answer *Moving*, one selected *None of the above*, and a later answer changed after the key was disclosed. The record does not support model rankings, hallucination scores, proprietary inference functions, or convergence theorems.

The proposed DRPAE framework converts this limitation into a methodological contribution. It separates key conformity from evidence compatibility, formalizes provenance, and adds robustness, calibration, expert adjudication, reproducibility, and uncertainty analysis. This approach permits educational systems to respect operational scoring while retaining scholarly scrutiny of item quality. More broadly, it replaces persuasive but non-identifiable stories about model reasoning with an auditable measurement design.

Declarations

Ethics approval: The study analyzed an author-supplied documentary artifact and did not recruit human participants or use identifiable private data. Institutional ethics review was therefore not applicable to this methodological reanalysis. Any future study involving student data, instructor judgments, or human-subject experiments should obtain the approvals required by the relevant jurisdiction and institution.

Consent for publication: Not applicable.

Data availability: The analyzed source artifact, the claim-provenance coding structure, and the equations used in this reanalysis are available from the corresponding author, subject to removal of any platform-identifying or

private instructional information. A confirmatory benchmark should release de-identified prompts, raw outputs, adjudication records, and analysis code where licensing and privacy permit.

Funding: No external funding was reported for this study.

Conflict of interest: Cuauhtémoc G. Gómez-D is affiliated with ARTIFEX Consulting, and Margarita García Nieto is affiliated with CECAPMKG Consulting. The authors declare no financial conflict directly related to the systems discussed in this manuscript.

Author contributions (CRediT): Cuauhtémoc G. Gómez-D: conceptualization, methodology, formal analysis, investigation, writing - original draft, visualization. Margarita García Nieto: validation, supervision, writing - review and editing. Both authors are responsible for reviewing and approving the final authorship statement before submission.

AI-assisted preparation disclosure: Generative AI tools were used for editorial organization, equation typesetting support, and language refinement. The authors retained responsibility for source selection, claim verification, theoretical adjudication, interpretation, and final approval. No AI system is listed as an author.

References

[1] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021) Measuring massive multitask language understanding. In: International Conference on Learning Representations.

[2] Pezeshkpour P, Hruschka E (2024) Large language models sensitivity to the order of options in multiple-choice questions. Findings of the Association for Computational Linguistics: NAACL 2024: 2006-2017. doi:10.18653/v1/2024.findings-naacl.130.

[3] Zheng C, Zhou H, Meng F, Zhou J, Huang M (2024) Large language models are not robust multiple choice selectors. In: International Conference on Learning Representations.

[4] Alzahrani N, Alyahya H, Alnumay Y, AlRashed S, Alsubaie S, Almushaykeh Y, Mirza F, Al-Otaibi N, AlTwairesh N, Alowisheq A, Nagoudi EMB, Abdul-Mageed M (2024) When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics 1: 13787-13805. doi:10.18653/v1/2024.acl-long.744.

[5] Wang X, Hu C, Ma B, Röttger P, Plank B (2024) Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In: Conference on Language Modeling 2024.

[6] Góral G, Wiśnios E, Sankowski P, Budzianowski P (2025) Wait, that's not an option: LLMs robustness with incorrect multiple-choice options. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics 1: 1495-1515. doi:10.18653/v1/2025.acl-long.75.

[7] Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36: 74952-74965.

[8] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) Standards for Educational and Psychological Testing. American Educational Research Association, Washington, DC.

[9] Messick S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 50(9): 741-749. doi:10.1037/0003-066X.50.9.741.

[10] Tam ZR, Wu CK, Lin CY, Chen YN (2025) None of the above, less of the right: Parallel patterns in human and LLM performance on multiple-choice question answering. Findings of the Association for Computational Linguistics: ACL 2025: 20112-20134. doi:10.18653/v1/2025.findings-acl.1031.

[11] Lewin K (1947) Frontiers in group dynamics: Concept, method and reality in social science; social equilibria and social change. Human Relations 1(1): 5-41. doi:10.1177/001872674700100103.

[12] Cummings S, Bridgman T, Brown KG (2016) Unfreezing change as three steps: Rethinking Kurt Lewin's legacy for change management. Human Relations 69(1): 33-60. doi:10.1177/0018726715577707.

[13] Kelman HC (1958) Compliance, identification, and internalization: Three processes of attitude change. Journal of Conflict Resolution 2(1): 51-60. doi:10.1177/002200275800200106.

[14] Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Amodei D, Kaplan J, Bowman SR (2023) Discovering language model behaviors with model-written evaluations. Findings of the Association for Computational Linguistics: ACL 2023: 13387-13434. doi:10.18653/v1/2023.findings-acl.847.

[15] Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1): 37-46. doi:10.1177/001316446002000104.