Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sycophantic AI feedback and disconfirming evidence search among business students

Pengfei Li

lipengfeizzgs@163.com

Zhengzhou Business University

Article

Keywords: AI sycophancy, evidence selection, business education, perceived judgement correctness, human–AI interaction

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11351646/v1

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

Additional Declarations: No competing interests reported.

Sycophantic AI feedback and disconfirming evidence search among business students

Pengfei Li

Zhengzhou Business University, China

Correspondence: lipengfeizzgs@163.com

Article

Abstract

Generative AI may influence learning by shaping which evidence students choose to encounter, even when it does not determine their final answer. This study examines whether sycophantic feedback is associated with reduced search for information opposing an initial business judgement and whether perceived judgement correctness accounts statistically for that relationship. A questionnaire-based decision task conducted in September 2025 yielded 405 retained behavioural records from students at 16 Chinese universities. Participants received one of two assigned feedback conditions and opened four of eight materials concerning a fictional business expansion decision. The sycophantic group allocated 36.7% of its openings to opposing materials, compared with 47.5% in the evidence-based comparison group, an unadjusted difference of −10.9 percentage points (95% confidence interval −15.5 to −6.2). An exploratory conditional choice model that respected the four-of-eight constraint and adjusted for material preferences within initial positions yielded an opposition-related odds multiplier of 0.458 (95% confidence interval 0.326–0.645). In the common 297-participant path-analysis sample, under the principal ordinary-least-squares model, the estimated indirect component through post-feedback correctness was −0.015 (bootstrap 95% confidence interval −0.031 to −0.002). This pathway was sensitive to school clustering and modest hypothetical residual confounding; the auxiliary self-report contrast was not significant. The findings locate a potential cost of affirming AI feedback at the evidence-acquisition stage. They support an account of altered information allocation more strongly than a uniquely identified psychological mechanism or a general decline in critical-thinking ability.

Keywords: AI sycophancy; evidence selection; business education; perceived judgement correctness; human–AI interaction

Introduction

Generative artificial intelligence increasingly participates in the formation of academic and professional judgements. Students can ask a conversational system to assess an argument, evaluate a business proposal or justify a preferred strategy before deciding what to read next. These interactions alter more than the efficiency of producing an answer. They can also alter the evidential basis on which that answer is developed. Reviews of educational AI identify opportunities for personalised support alongside concerns about dependence, inaccurate outputs and inadequate evaluation of learning processes (Kasneci et al., 2023; Lo, 2023; Yan et al., 2024). An important question is therefore whether assistance preserves the learner's willingness to expose an initial conclusion to challenge.

Evidence about the benefits of generative AI does not settle this question. AI assistance improved professional writing performance in Noy and Zhang (2023), increased individual creative evaluations while reducing collective diversity in Doshi and Hauser (2024), and improved immediate learning outcomes in a particular tutoring intervention in Kestin et al. (2025). A meta-analysis found that the performance of human–AI combinations varies substantially across tasks and comparison benchmarks (Vaccaro et al., 2024). These findings make the evaluation target consequential. A well-written product, a correct final response and a process that actively considers contrary evidence describe different achievements. An intervention can improve one without establishing improvement in the others.

Sycophancy creates a specific concern at this process level. Sharma et al. (2024) describe language-model behaviour that favours agreement with users' expressed beliefs over truthfulness and show that users' stated preferences can influence feedback on the same material. In an educational interaction, such feedback may do more than communicate approval. It can make a provisional judgement appear sufficiently justified to reduce the perceived value of further scrutiny. This possibility connects output-level research on agreeable model responses with the downstream behaviour of the person receiving them. Whether the user then seeks challenging information is an empirical question; it cannot be inferred from the wording of the response alone.

The present study examines this question through disconfirming evidence search: the proportion of a fixed set of opened materials that opposes the participant's initial business recommendation. This is an observable allocation of information opportunities. It differs from self-reported critical thinking and from the objective accuracy of a final answer. The distinction matters because subjective evaluations and proxy tasks can produce different conclusions from behavioural measures in human–AI research (Buçinca et al., 2020). Lee et al. (2025), for example, connect confidence in AI and task-specific self-confidence to self-reported critical-thinking effort among knowledge workers. Their findings motivate closer attention to confidence and effort, but do not establish how experimentally varied affirmation affects actual evidence selection.

Business education provides a theoretically useful setting because strategic choices often involve defensible but competing criteria. In the present task, students recommend either continued direct operation or expansion through franchising for a fictional tea brand. Control and consistency favour one approach; scale and capital efficiency favour the other. Neither criterion establishes a uniquely correct recommendation without further assumptions about implementation and market conditions. An opposing material can therefore have value by revealing an overlooked constraint, even when it does not overturn the chosen strategy. The task permits an examination of exposure to challenge without equating disagreement with error.

The analysis distinguishes assigned feedback, perceived feedback characteristics, subjective correctness and recorded information choice. It first estimates whether the feedback conditions differ in the share of opposing materials opened. It then asks whether post-feedback perceived judgement correctness contributes a statistical indirect pathway after adjustment for baseline correctness. A

further choice model evaluates the condition contrast while accounting for the fixed four-of-eight selection rule and preferences for individual materials. These analyses address different questions: the aggregate contrast describes how much allocation changes, the choice model examines whether the pattern persists after accounting for material preferences, and the path model evaluates a candidate psychological explanation.

The contribution is to locate an observable consequence of AI feedback between receiving advice and evaluating a final product. This perspective extends discussion of appropriate reliance from accepting or rejecting advice to obtaining the information needed to assess it. It also gives a bounded behavioural application to the socioaffective alignment agenda advanced by Kirk et al. (2025): satisfying a user's immediate preference for validation may not support the longer-term goal of revisable judgement. The study concerns a single decision episode, rather than sustained attachment to AI. Its theoretical value lies in specifying how approval might reshape the information available for subsequent reasoning.

Conceptual framework and hypotheses

Sycophancy as a feature of evaluative feedback

Sycophantic feedback is operationalised here as an evaluation that privileges agreement with a user's initial judgement over proportionate scrutiny of its evidential basis. The distinguishing feature is the relationship between endorsement and justification. Encouragement is not necessarily sycophantic, and agreement can be warranted. Feedback becomes problematic when its affirmation tracks the user's preferred answer more strongly than the evidence warrants. This operational definition follows the feedback-related concern examined by Sharma et al. (2024), while specifying the exposure studied here: a presented evaluation of an already expressed business recommendation.

Persuasiveness and epistemic value can diverge in human–AI interaction. Bansal et al. (2021) found that explanations increased acceptance of AI recommendations without consistently improving complementary team performance. Poursabzi-Sangdeh et al. (2021) similarly showed that greater interpretability did not reliably help participants correct model errors. Explainability can affect perceived understanding, trust and acceptance (Shin, 2021), whereas the broader trust literature distinguishes cognitive and emotional bases of reliance on AI (Glikson and Woolley, 2020). Together, these studies motivate a distinction between feedback that is readily accepted and feedback that supports selective evaluation. Acceptance alone provides an incomplete account of educational value.

Sycophancy is particularly relevant after a learner has committed to a position. At that point, feedback can act as apparent corroboration: the system seems to endorse a judgement that is already personally salient. An approving evaluation may be treated as additional support even when it contains little independent evidence. The proposed account does not require students to regard the system as infallible. It requires only that endorsement changes the relative usefulness they assign to examining alternative considerations. The downstream consequence should then appear in what they choose to inspect, rather than necessarily in their reported satisfaction with the interaction.

Allocation of a fixed opportunity to search

The study treats evidence acquisition as a constrained allocation problem. Each participant opens four of eight materials, so the total number of openings is held constant. A lower proportion of opposing materials consequently implies greater allocation to materials consistent with the initial position. It does not imply less total clicking, a shorter overall search or a decline in general cognitive ability. This constraint makes the outcome interpretable: feedback may change the composition of acquired evidence even when the interface offers the same quantity of access.

The proposed process concerns the perceived marginal value of another search opportunity. If feedback presents the existing recommendation as well founded, an opposing source may seem less

likely to improve the decision. The student can instead use a limited opening to elaborate or defend the preferred proposal. Feedback that identifies an unresolved trade-off gives challenge a more immediate purpose. A student recommending direct operation might investigate whether control is worth slower market coverage; one recommending franchising might investigate whether scale offsets execution risk. Under this account, counterevidence becomes attractive when it can resolve uncertainty that remains relevant to the choice.

This argument does not assume that balanced exposure is always optimal. A contrary source can be weak, redundant or irrelevant, and the correct allocation depends on its evidential value. The task's outcome measures willingness to encounter the opposing position under a specified coding scheme. It does not award epistemic superiority to every opposing passage. This qualification is conceptually important: search is one component of critical evaluation, which also requires assessing source quality and integrating reasons. The present model focuses on the opportunity for those later operations to occur.

The feedback contrast combines affirmation with differences in argumentative balance. For students initially choosing direct operation, the sycophantic text endorses their strategic direction, whereas the comparison text additionally raises growth and market coverage. For students initially choosing franchising, the comparison text introduces quality control and store survival. Thus, the exposure is a pair of feedback packages. One consolidates reasons for the chosen option; the other makes competing criteria explicit. Both the reduced need to reconsider and the salience of opposing reasons predict a difference in subsequent evidence allocation.

H1. Participants assigned to sycophantic feedback will open a lower proportion of materials opposing their initial judgement than participants assigned to evidence-based feedback.

Perceived judgement correctness as a candidate pathway

Perceived judgement correctness is the subjective sense that one's evaluation is valid or right. It differs from attitude clarity, which concerns knowing what one thinks. Blankenship et al. (2022) examine this distinction and discuss external validation as relevant to correctness. Applied to the present task, affirming feedback may be experienced as corroboration of the initial recommendation. The construct concerns a belief about a particular judgement; without an objective benchmark, a high score cannot establish accuracy or overconfidence.

The object of confidence is central to the model. Trust in AI concerns the adviser, whereas self-efficacy concerns the learner's capability. Perceived correctness concerns the judgement already made. A learner can trust an AI system while doubting a particular recommendation, or feel capable of reasoning while recognising that the present conclusion may be wrong. Consequently, confidence in an initial answer should not be treated as interchangeable with confidence in one's ability to evaluate new evidence. Keeping these referents distinct makes the proposed mechanism more precise than a general claim that confidence impairs critical thinking.

Baseline correctness, M_0, allows the analysis to distinguish prior conviction from the post-feedback state, M_1. A positive relation between assignment and M_1 after conditioning on M_0 would be consistent with feedback supplying additional subjective validation. It would not mean that every initially confident student became more confident, nor would it establish how the student interpreted the source of that validation. Perceived sycophancy, P, is retained as a manipulation-check measure. It is not substituted for assigned condition, and it does not enter the mediation equations as an additional predictor.

H2. Sycophantic feedback will predict higher post-feedback perceived judgement correctness after adjustment for baseline correctness and the stated pre-feedback covariates.

There are competing reasons why correctness might relate to subsequent search. Rollwage et al. (2020) found that experimentally increased confidence biased later evidence accumulation towards an initial decision in a perceptual task. Schulz et al. (2020) linked dogmatism to reduced information

search, particularly under uncertainty. These findings support a possible route from the sense of already being right to reduced demand for challenge. However, their constructs and tasks do not identify the same process as correctness judgements about a business strategy. Rollwage and Fleming (2021) also show that confirmatory processing can be adaptive under particular metacognitive conditions.

A competing expectation is that confidence can sustain engagement. Lee et al. (2025) report that greater self-confidence in task performance was associated with greater self-reported critical-thinking engagement, whereas confidence in AI showed a different association. A student who feels able to defend a proposal may be more willing to confront disagreement. The model therefore retains a non-directional hypothesis for the indirect component. It tests whether the post-feedback correctness pathway is distinguishable from zero without treating every confidence-related process as necessarily harmful.

H3. Feedback condition will have a non-zero statistical indirect association with disconfirming evidence search through post-feedback perceived judgement correctness, conditional on baseline correctness and the stated covariates.

Two pathways can coexist. Feedback may change how correct the student feels, and it may directly cue an alternative consideration that becomes easier to retrieve during selection. The remaining condition coefficient is compatible with several such processes; it is not a measured counterargument-salience effect. Nor does conditioning on M_1 establish that the indirect component is causal. The separation between statistical decomposition and psychological identification follows the concerns about popular path models discussed by Rohrer et al. (2022) and the call to state causal assumptions explicitly in Grosz et al. (2020). Figure 1 shows the hypothesised structure. These are theory-guided hypotheses evaluated retrospectively; a prospective preregistration is not claimed.

Methods

Participants and design

Recruitment was conducted through course announcements and student academic networks at the participating institutions between 3 and 30 September 2025. Participation was voluntary, was unrelated to course grades and carried no monetary incentive. Students enrolled in business-related undergraduate or postgraduate programmes were eligible to take part.

The completion codes identify 15 incomplete behavioural tasks, 15 duplicate-submission flags, 10 technical failures and five attention-check failures. Retaining the remaining records reproduces the 405-record analysis file exactly, with no changes to retained data cells. The submitted records comprise 228 assignments to $X = 0$ and 222 to $X = 1$; the retained groups contain 204 and 201 participants. Figure 2 and Supplementary Table S2 show the group-specific exclusions. These are recorded classifications.

Duplicate submissions were removed by matching device identifiers and completion codes, and inattentive respondents were identified through two instructed-response attention checks embedded in the survey.

The sample size was set by the number of students recruited during the data-collection window; following Lakens (2022), the analysis prioritises precise estimation and reports effect sizes with confidence intervals rather than relying on a post hoc power argument. Age and gender were not collected to reduce response burden.

Business task and feedback

The case concerned the fictional fresh-tea brand Chaxu (茶序). The scenario described 80 directly operated stores, 62 located in Jiangsu, Zhejiang and Anhui, a typical transaction value of RMB18–22,

and an established supply chain and membership system. These are fictional case premises, rather than observations about an actual company. Option A proposed continued direct operation and expansion to approximately 200 stores over three years. Option B proposed regional franchising and expansion to approximately 500 stores. The decision involved competing priorities of control, consistency, growth and capital efficiency; the task did not specify an objectively correct option.

Participants first selected A or B and completed M_0. The study sequence then placed feedback X before M_1, followed by four material selections/openings, the auxiliary ratings and a final A/B decision. The archive contains four feedback variants, crossing assigned condition with initial position. Each variant addresses the proposal the student had already selected. The sycophantic texts strongly endorse that proposal; the evidence-based texts acknowledge its rationale while explicitly introducing considerations supporting the alternative. The comparison therefore engages the initial judgement in both conditions while varying endorsement and argumentative balance together.

After the baseline responses, the survey platform randomly assigned participants to one of two feedback conditions with probability 0.50, then displayed the fixed feedback text corresponding to the assigned condition and the participant's initial A/B choice. Feedback was delivered as a pre-written, AI-labelled message rather than a live model-generated conversation. The eight supporting materials were presented under neutral identifiers in a fixed order, and the interface allowed each participant to open exactly four of them.

The eight materials comprise four passages favouring direct operation, m1–m4, and four favouring franchising, m5–m8. An opposing opening is coded relative to the initial position, even if the final choice changes. Thus m5–m8 oppose A, whereas m1–m4 oppose B. This mapping exactly reproduces the recorded opposing-material count in all 405 retained observations. The full Chinese case, materials and feedback, with English reporting translations, appear in Supplementary S0. The translations were not a separately administered instrument.

The supplementary archive distinguishes the supplied stimulus wording from subsequent source checks. The text lengths are 152–207 non-whitespace characters for the materials and 175–205 for the feedback variants, counting punctuation and excluding headings and source lines. Perceived sycophancy and perceived information quality were measured within the main experiment as manipulation checks. The analysis therefore treats the perceived contrast as empirical evidence of the manipulation rather than assuming equivalence of perceived information quality or feedback intensity across conditions.

Measures and scoring

The primary outcome is disconfirming evidence search, $Y = D/4$, where D is the number of opened materials opposing the initial position. Y takes values 0, 0.25, 0.50, 0.75 and 1. Every retained participant has four distinct opened IDs, and the SELECT and OPEN fields agree throughout the retained data. Opening is an observable information choice. The outcome does not incorporate reading duration, comprehension or how a passage is subsequently used in reasoning. The final A/B choice is retained as a separate field and does not redefine which openings count as disconfirming.

Perceived judgement correctness is assessed by three nine-point items before and after feedback. They concern certainty that the judgement is correct, that others with the same information should reach the same judgement, and that the judgement is the most reasonable of the available alternatives. The post-feedback items refer to the initial judgement. M_0 and M_1 are the corresponding item means. Blankenship et al. (2022) provide a recent conceptual and measurement discussion relevant to correctness; their paper is not represented as the original source or validation of the exact Chinese wording administered here. The complete wording is reported in Supplementary S3.

Perceived sycophancy, P, comprises four seven-point items addressing endorsement ahead of scrutiny, affirmation despite insufficient evidence, endorsement exceeding the available evidence and avoidance of necessary questioning. These are study-specific manipulation-check items. Self-reported

disconfirming search, Y_s, comprises four seven-point items addressing potentially overturning information, inconsistent evidence, limitations of the initial proposal and reasons supporting the alternative. Y_s is an auxiliary measure and is not treated as interchangeable with recorded Y. Both scales require validation beyond the internal measurement evidence provided by this sample.

The pre-feedback covariate vector C includes prior coursework directly relevant to the case (0 = no, 1 = yes), the number of days using generative AI in the preceding seven days (0–7), and prior familiarity with this type of decision (1–7). Baseline correctness is entered separately. These variables index relevant prior experience; they are not measures of AI literacy, academic achievement or general critical-thinking ability. Neither P nor Y_s is used to exclude observations or to define a post-treatment adjustment set.

Primary composite scores require all items in a block to be observed. This yields M_0 for 354 participants, M_1 for 340, P for 325 and Y_s for 342. The common complete M_0/M_1 sample contains 297 participants and is used for all three path equations. The source workbook also reports available-item means, which retain all 405 participants. Analyses using those means are reported as scoring sensitivities rather than multiple imputation. This explicit distinction follows the measurement-transparency concerns raised by Flake and Fried (2020). Missingness and sample denominators are reported for each analysis.

Primary models and uncertainty

The unadjusted behavioural comparison uses all 405 retained records. The baseline-adjusted total contrast uses the 354 records with complete M_0 and observed covariates. Path decomposition uses the common 297-participant sample so that its total, direct and indirect components refer to the same observations and adjustment set. The linear equations are:

$$Y_i = \tau_0 + cX_i + \rho_0 M_{0i} + \theta^T C_i + \varepsilon_i \quad (1)$$

$$M_{1i} = \alpha_0 + aX_i + \rho_1 M_{0i} + \gamma^T C_i + u_i \quad (2)$$

$$Y_i = \beta_0 + c'X_i + bM_{1i} + \rho_2 M_{0i} + \delta^T C_i + v_i \quad (3)$$

$$\mathrm{IE} = ab, \quad \mathrm{DE} = c', \quad \mathrm{TE} = c = c' + ab \quad (4)$$

Here c is the adjusted total condition contrast; a relates condition to post-feedback correctness; b relates M_1 to Y after adjustment; and c′ is the remaining condition coefficient. IE, DE and TE denote statistical indirect, direct and total components. Coefficients are unstandardised. The specification contains no condition-by-mediator interaction. The identity in equation (4) holds for these fitted linear models on a common sample, but does not identify a causal mediator. Randomisation of X, even when implemented as reported, does not randomise M_1 or remove common causes of M_1 and Y.

Ordinary least squares is used to estimate interpretable mean differences. Multiplying a coefficient in a Y model by 100 expresses percentage points. Principal confidence intervals use heteroskedasticity-robust HC3 standard errors with a normal reference. The indirect component uses 5,000 participant-level bootstrap resamples, percentile intervals and seed 20250930. An unadjusted two-sided permutation comparison uses 19,999 reallocations preserving retained group sizes, seed 20250903 and the plus-one correction; its interpretation assumes exchangeable individual assignment in the retained sample. Additional analyses use school-clustered standard errors with a finite-sample correction and t(15), school fixed effects, and an ordered logit retaining all five D levels.

Measurement assessment reports Cronbach's alpha and a correlated-factor confirmatory model for M_0, M_1, P and Y_s among 195 participants with all 14 items observed. The model permits residual covariances between each repeated correctness item before and after feedback and treats the seven- and nine-point responses as continuous. Standardised composite omega and average variance extracted are computed from the fitted loadings and residual variances, with interpretation informed by Hayes and Coutts (2020) and Flora (2020). Absolute-correlation HTMT summaries assess the distinct post-

feedback constructs. M_0 and M_1 are repeated measurements of one construct rather than competing constructs required to demonstrate discriminant validity. Measurement conclusions are sample-specific, consistent with Rönkkö and Cho (2022).

Conditional choice and mechanism sensitivity

An exploratory conditional choice analysis addresses a feature that the aggregate Y model leaves implicit: selecting one material reduces the remaining opportunities to select another. Let S_i be the four-element set opened by participant i and A_4 the 70 possible four-element subsets of the eight materials. For initial position g_i, material j has a position-specific parameter η_{jg}. Let O_{ij} indicate that material j opposes participant i's initial choice. The probability assigned to a selected set is:

$$\Pr(S_i \mid |S_i| = 4, X_i, g_i) = \frac{\exp(\sum_{j \in S_i} w_{ij})}{\sum_{A \in A_4} \exp(\sum_{j \in A} w_{ij})}$$
$$w_{ij} = \eta_{jg_i} + \lambda X_i O_{ij} \quad (5)$$

The model conditions on each participant selecting exactly four materials. Position-specific material parameters allow, for example, m2 to have a different baseline appeal to students initially choosing A and B; m8 is the reference in each position. The coefficient λ captures the additional relative preference for opposing rather than supporting materials in the sycophantic condition. Its exponential is a conditional odds multiplier for exchanging a supporting item for an opposing one, comparing feedback conditions while holding the other selected items and initial position fixed. It is not a percentage-point effect or a marginal probability ratio.

The likelihood is evaluated by enumerating all 70 sets for each of the four initial-position/condition combinations. Participant-level score contributions provide robust uncertainty, with an additional school-clustered calculation using 16 clusters and t(15). All 405 participants enter the model; the 3,240 participant–material alternatives are not counted as independent participants. An independent conditional-logit implementation verifies the likelihood and parameter estimates. Supplementary S10 supplies the full 15-parameter results and observed item-selection counts. This analysis was added during revision and is identified as exploratory.

A separate sensitivity calculation asks how the estimated indirect component changes when the structural errors of the linear mediator and outcome equations have a hypothetical correlation ρ. The calculation adjusts the estimated b coefficient under that linear structural-error specification and identifies the value of ρ at which ab becomes zero. The observed data do not identify ρ. The resulting curve is a deterministic sensitivity of point estimates, not a confidence band, a simulated dataset or evidence that a particular omitted variable exists. The derivation and numerical grid appear in Supplementary S11 and Figure 5.

The added evaluation workbook is reconciled to the individual records before use. Its item and available-item composite summaries correspond to equal-variance two-sample t tests, with the reported t statistic oriented X = 0 minus X = 1. The supplementary item comparisons consistently use X = 1 minus X = 0, report Welch inference and apply Holm adjustment across 14 exploratory tests. Initial-position differences are assessed by an X-by-position interaction, rather than by comparing subgroup significance labels. Primary tests and intervals are two-sided and unadjusted for multiplicity; auxiliary tests are explicitly distinguished.

Analyses were conducted in Python using NumPy, pandas, SciPy, statsmodels and semopy. Exact versions, source-file hashes, seeds and executable code accompany the review package. Generative AI assistance through OpenAI Codex supported literature searching, manuscript drafting, analysis-code development and figure preparation. The human author remains responsible for checking the records, sources, analytical choices and final text. The author confirms that all numerical results reported in this manuscript are derived directly from the retained analysis dataset.

Results

Sample composition and measurement

The retained sample includes 107 participants from Henan, 104 from Jiangsu, 101 from Shandong and 93 from Guangdong. There are 361 undergraduates and 44 postgraduates; 234 participants report prior relevant coursework. Complete M_0 is available for 181 comparison-condition and 173 sycophantic-condition participants. The common path-analysis sample comprises 148 and 149 participants, respectively. Table 1 and Supplementary S1 give the full composition. The different analytic denominators reflect missing questionnaire items, rather than the removal of participants based on observed effect direction.

Internal consistency estimates are 0.883 for M_0, 0.915 for M_1, 0.891 for P and 0.843 for Y_s (Table 2). The four-factor model yields $\chi^2(68) = 65.470$, $p = 0.564$, RMSEA = 0.000, and CFI and TLI rounded to 1.000 after bounding at unity. Standardised loadings range from 0.731 to 0.919, standardised omega from 0.834 to 0.919, and average variance extracted from 0.557 to 0.791. HTMT is 0.089 for M_1–P, 0.254 for M_1–Y_s and 0.071 for P–Y_s. Supplementary Tables S4–S5 give the unbounded fit indices and item estimates. These results describe the internal structure of the observed responses; they do not establish general validation of the newly specified items.

Perceived sycophancy is higher in the sycophantic condition: means are 4.823 and 3.633 among 325 complete P records. The unadjusted difference is 1.190 points (95% CI 0.956 to 1.425, $p < 0.001$). Participants therefore perceived a difference in the feedback characteristics targeted by P. This comparison does not demonstrate that the conditions were equivalent on all other attributes, particularly the prominence of reasons supporting the alternative.

Behavioural condition contrast

Participants in the comparison condition open an average of 1.902 opposing materials out of four, compared with 1.468 in the sycophantic condition. Mean Y is 0.475 (SD 0.247) and 0.367 (SD 0.233), respectively. The unadjusted difference is −0.1086 (95% CI −0.1555 to −0.0617, $p < 0.001$), equivalent to 10.9 percentage points or approximately 0.434 fewer opposing openings per participant. The pooled-standard-deviation effect is $d = -0.452$, and the permutation p value is 0.0001. Figure 3 and Table 3 report the full distribution.

The contrast is visible across the outcome categories. No opposing material is opened by 16 of 204 comparison participants (7.8%) and 31 of 201 sycophantic participants (15.4%). Three or four are opened by 58 comparison participants (28.4%) and 27 sycophantic participants (13.4%). Adjustment for M_0 and the three prior-experience covariates yields a total condition contrast of −0.1256 among 354 participants (95% CI −0.1748 to −0.0763, $p < 0.001$). Both the unadjusted and adjusted estimates support H1 in the retained observations.

Correctness and indirect association

On the common 297-participant sample, sycophantic assignment predicts higher post-feedback correctness after adjustment: $a = 0.5797$ (95% CI 0.3173 to 0.8421, $p < 0.001$), supporting H2. Higher M_1 is associated with lower Y in the outcome equation, $b = -0.0258$ (95% CI −0.0488 to −0.0029, $p = 0.027$). The total contrast is $c = -0.1013$, and the remaining direct component is $c' = -0.0864$ (95% CI −0.1424 to −0.0303, $p = 0.003$). Table 4 reports all coefficients, including the covariates, for the same sample.

The statistical indirect component is $ab = -0.0150$, with a 5,000-resample percentile interval of −0.0315 to −0.0015. This meets the non-zero statistical criterion in H3 under the principal model. In outcome units, the component is approximately 1.5 percentage points of the search proportion. Its magnitude is appreciably smaller than the remaining condition contrast. The decomposition does not

establish how much of a causal effect is mediated, because the mediator–outcome relation requires identifying assumptions beyond assignment of feedback. Taken together, this component should be read as a model-dependent exploratory estimate rather than a confirmed mediation effect; the behavioural contrast, not the pathway, is the result that survives the sensitivity checks.

Robustness of allocation and fragility of the mechanism

The common-sample total contrast remains negative with school-clustered inference (−0.1013, 95% CI −0.1612 to −0.0415, p = 0.003) and school fixed effects (−0.1044, 95% CI −0.1599 to −0.0489, p < 0.001). Available-item scoring in all 405 records produces a total contrast of −0.1155 and an indirect component of −0.0159. The ordered-logit outcome model retains a negative condition coefficient (−0.653, p = 0.003; odds ratio 0.520) under its proportional-odds specification. Figure 4 and Table 5 compare these analyses. They reuse overlapping records and are sensitivity checks, not independent replications.

The conditional set-choice model yields $\lambda = -0.7805$ and an opposition-related odds multiplier of 0.458 (participant-robust 95% CI 0.326 to 0.645, p < 0.001). The school-clustered interval is 0.332 to 0.632 (p < 0.001). Thus, allowing each material to have a position-specific baseline preference and respecting the constrained selection set does not remove the negative allocation pattern. This odds multiplier concerns the relative selection of opposing versus supporting materials and should not be described as a 54.2% reduction in the probability of any individual opening. Table 6 and Supplementary S10 report the model.

Mechanism evidence is more sensitive. With school-clustered standard errors, b remains −0.0258 but its interval spans −0.0533 to 0.0016 (p = 0.063). The structural-error sensitivity places the zero of the indirect point estimate at $\rho = -0.123$ (Figure 5). Under that stipulated linear specification, a negative residual association of that size is sufficient to account for the estimated b pathway without a non-zero structural mediator coefficient. This does not estimate actual confounding. It shows that the proposed mechanism is less secure than the observed condition contrast, even though the principal bootstrap interval excludes zero.

Exploratory initial-position estimates are −0.1410 for position A (n = 219, 95% CI −0.2028 to −0.0792) and −0.0711 for position B (n = 186, 95% CI −0.1428 to 0.0006). The interaction is 0.0699 (95% CI −0.0247 to 0.1645, p = 0.148). Both point estimates are negative, but the data establish neither a difference between branches nor equivalence. The distinct subgroup p values do not constitute a test of differential effects. These results are reported in Supplementary S9.

The auxiliary self-report contrast is not statistically significant. Y_s means are 4.199 and 3.997, with a difference of −0.201 (95% CI −0.424 to 0.021, p = 0.076; n = 342). Its correlation with recorded Y is 0.331. The available-item comparison uses 405 records and gives pooled-variance p = 0.0529 and Welch p = 0.0531; both remain above 0.05. YS2 alone has an unadjusted Welch p = 0.0160, but its Holm-adjusted value across the 14 item tests is 0.1122. An isolated item result does not establish a significant scale-level outcome. Supplementary S8 reconciles these scoring and inferential differences.

Discussion

Evidence acquisition as a consequence of AI feedback

The principal finding is a shift in the composition of the evidence students choose to encounter. Given four openings, students receiving sycophantic feedback allocate fewer to materials opposing their initial recommendation. The average difference is substantial enough to be behaviourally interpretable, yet it describes a narrow outcome: approximately 0.43 of the four openings. The contribution is therefore not that AI universally weakens critical thinking. It is that evaluative feedback can be associated with an earlier stage at which the opportunity for critical evaluation is distributed across supporting and challenging information.

This stage is theoretically consequential because later reasoning operates on an acquired evidence set. A polished answer can conceal a restricted set of considerations, while a less polished answer may reflect serious engagement with alternatives. Fan et al. (2025) show that different forms of learning support can change learning processes and performance in ways not captured by a single general benefit claim. Zhai et al. (2024) organise concerns about over-reliance across several cognitive capabilities. The present result adds a specific behavioural target: whether assistance preserves access to information capable of placing pressure on the learner's own assumptions.

The finding extends the appropriate-reliance perspective without redefining it. Schemmer et al. (2023) distinguish accepting correct advice from rejecting incorrect advice. In strategic decisions, the correctness of advice is often not directly observable at the moment of use. Evidence acquisition then helps determine which reasons deserve reliance. In the Chaxu case, searching for execution risks can be productive even if franchising remains the preferred strategy; examining growth opportunities can be productive even if direct operation is retained. The epistemic role of challenge is to test the conditions under which a recommendation holds, rather than to compel a reversal for its own sake.

The conditional choice analysis sharpens this interpretation. Some passages may attract attention because of their topic, numerical detail or familiar company examples. A difference in aggregate opposing openings could therefore be entangled with baseline material preferences. Allowing those preferences to differ by initial stance retains the condition-related opposition contrast. The selection constraint is also modelled directly, so the result does not rely on treating four openings as four independent trials. This convergence strengthens the allocation finding while leaving the content of the feedback contrast itself unchanged.

Distinguishing conviction from epistemic competence

The correctness pathway helps specify a candidate process. Feedback can offer apparent corroboration of an already expressed recommendation, and the student may then see less need to inspect information that could qualify it. The positive a coefficient and negative indirect component are consistent with this account. However, the measure concerns conviction that a particular judgement is right. It does not measure the ability to reason well, the accuracy of the judgement, or a general preference to avoid cognitive effort. These distinctions prevent a broad claim about confidence from being inferred from a narrowly defined subjective state.

They also help reconcile the proposed account with apparently different findings in the literature. The self-confidence measured by Lee et al. (2025) concerns confidence in task performance and can accompany stronger reported critical engagement. Rollwage et al. (2020) examine confidence in an initial perceptual decision, which can bias subsequent accumulation. The relevant theoretical question is thus what the person is confident about and how that confidence changes the expected value of further information. Confidence in the capacity to evaluate objections may encourage search; confidence that the present conclusion already needs little revision may reduce it. The current data assess the second referent without directly comparing the two.

The mechanism sensitivity prevents this explanation from becoming stronger than the design allows. The principal model yields a non-zero indirect component, but school clustering weakens the b-path evidence and a modest hypothetical residual correlation removes its point estimate under the linear sensitivity specification. These findings give different evidential weight to the behavioural contrast and the psychological pathway. The former is supported across several representations of the observed choices. The latter remains a plausible process requiring a design that independently varies or more securely identifies the mediator. Path decomposition alone cannot supply that identification (Rohrer et al., 2022). Accordingly, the present evidence supports a claim about what students choose to read after sycophantic feedback; on the strength of these analyses alone it does not establish why they do so.

Counterargument salience as a competing explanation

The feedback texts suggest another route. Evidence-based feedback explicitly names reasons to inspect the alternative; sycophantic feedback consolidates the preferred proposal. These cues may change which material seems relevant even if the student's subjective correctness remains unchanged. The two routes are not mutually exclusive. Affirmation can reduce uncertainty while an explicit counterargument increases the accessibility of a neglected criterion. Their coexistence explains why the manipulation can produce a clear behavioural difference even when the estimated correctness component is relatively small.

The remaining coefficient c' should not be renamed a counterargument-salience effect. Salience was neither independently manipulated nor directly measured. Instead, the texts identify a concrete alternative explanation that a subsequent factorial design could distinguish: vary endorsement strength while holding the reasons constant, and vary the presence of counterarguments while holding endorsement constant. Correctness, perceived information value and actual search could then be assessed in temporal order. This prediction makes the theoretical account testable without retrospectively inserting an unmeasured variable into the present model.

The divergence between recorded and self-reported search is also informative. The positive correlation between Y and Y_s indicates shared information, but their condition contrasts do not yield identical evidence. A student may regard an information search as sufficiently critical while opening relatively few opposing materials, or may interpret the self-report items through the quality of reasoning rather than the number of opposing openings. This is consistent with Buçinca et al. (2020), who caution that subjective and behavioural evaluations can address different aspects of human–AI performance. Retaining both outcomes reveals that distinction rather than obscuring it in a single general critical-thinking label.

Implications for learning and AI evaluation

For business teaching, the results motivate assessment of the evidential route to a recommendation. Instructors could require students to identify a condition under which their preferred proposal would fail, inspect a relevant opposing source and explain whether that source changes their reasoning. Such an exercise targets how a recommendation is justified rather than rewarding disagreement mechanically. The present study does not test this intervention, but its design logic is consistent with evidence that cognitive-forcing interventions can reduce over-reliance in AI-assisted decisions (Buçinca et al., 2021).

For AI feedback design, the findings support separating evaluative confidence from the provision of reasons. Zhang et al. (2020) examine confidence and explanation cues in trust calibration, showing why these cues deserve treatment as design variables. A feedback interface could qualify endorsement by specifying unresolved criteria and the evidence that would warrant revision. AI literacy includes critically evaluating technologies and using them in informed collaboration (Long and Magerko, 2020). In a business curriculum, that competence can be assessed through source evaluation, consideration of a strong objection and explicit conditions for revising a decision.

Source quality remains essential. Human heuristics for judging AI-generated language are imperfect (Jakesch et al., 2023), and generated language can make both accurate information and misinformation persuasive (Spitale et al., 2023). Pennycook and Rand (2021) likewise emphasise attention to accuracy in information judgement. An instruction to seek counterevidence should therefore include a requirement to assess its credibility and relevance. More opposing openings are educationally valuable when they create a better opportunity to evaluate assumptions, not simply because they increase a count.

For system evaluation, the broader implication is to connect feedback characteristics with subsequent user behaviour. Kirk et al. (2025) call attention to the social and psychological consequences of interactions that evolve between users and AI. The present task offers a limited but

measurable instance: immediate validation and continued scrutiny need not move together. Evaluation can therefore distinguish the feedback delivered, the learner's sense of correctness, the evidence selected and the quality of eventual reasoning. The current evidence concerns the middle stages and identifies a concrete question for future work on longer interactions.

Conclusion

This study tested whether affirming, sycophantic feedback from a generative system changes which evidence business students choose to inspect. Across 405 students at 16 universities, those who received sycophantic feedback allocated roughly ten fewer percentage points of their fixed information openings to materials that opposed their initial recommendation, and the effect held across alternative model specifications. The estimated pathway through post-feedback correctness was small and sensitive to school clustering and residual-correlation assumptions, so the result is best read as a behavioural consequence of AI feedback rather than as evidence for a settled psychological mechanism. As generative systems move from answering questions to shaping how learners justify their judgements, tracking what people read after the feedback—rather than only what they finally conclude—offers a practical point of scrutiny for both teaching and system design.

References

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems:1–16. https://doi.org/10.1145/3411764.3445717

Blankenship KL, Kane KA, Machacek MG (2022) Values and attitude certainty: The case for attitude clarity and correctness. Frontiers in Psychology 13:975864. https://doi.org/10.3389/fpsyg.2022.975864

Buçinca Z, Lin P, Gajos KZ, Glassman EL (2020) Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. Proceedings of the 25th International Conference on Intelligent User Interfaces:454–464. https://doi.org/10.1145/3377325.3377498

Buçinca Z, Malaya MB, Gajos KZ (2021) To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1):188. https://doi.org/10.1145/3449287

Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances 10(28):eadn5290. https://doi.org/10.1126/sciadv.adn5290

Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, Shen Y, Li X, Gašević D (2025) Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology 56(2):489–530. https://doi.org/10.1111/bjet.13544

Flake JK, Fried EI (2020) Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Advances in Methods and Practices in Psychological Science 3(4):456–465. https://doi.org/10.1177/2515245920952393

Flora DB (2020) Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. Advances in Methods and Practices in Psychological Science 3(4):484–501. https://doi.org/10.1177/2515245920951747

Glikson E, Woolley AW (2020) Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals 14(2):627–660. https://doi.org/10.5465/annals.2018.0057

Grosz MP, Rohrer JM, Thoemmes F (2020) The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. Perspectives on Psychological Science 15(5):1243–1255. https://doi.org/10.1177/1745691620921521

Hayes AF, Coutts JJ (2020) Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. Communication Methods and Measures 14(1):1–24. https://doi.org/10.1080/19312458.2020.1718629

Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences 120(11):e2208839120. https://doi.org/10.1073/pnas.2208839120

Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103:102274. https://doi.org/10.1016/j.lindif.2023.102274

Kestin G, Miller K, Klales A, Milbourne T, Ponti G (2025) AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports 15(1):17458. https://doi.org/10.1038/s41598-025-97652-6

Kirk HR, Gabriel I, Summerfield C, Vidgen B, Hale SA (2025) Why human–AI relationships need socioaffective alignment. Humanities and Social Sciences Communications 12(1):728. https://doi.org/10.1057/s41599-025-04532-5

Lakens D (2022) Sample Size Justification. Collabra: Psychology 8(1):33267. https://doi.org/10.1525/collabra.33267

Lee HP, Sarkar A, Tankelevitch L, Drosos I, Rintel S, Banks R, Wilson N (2025) The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems:1–22. https://doi.org/10.1145/3706598.3713778

Lo CK (2023) What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences 13(4):410. https://doi.org/10.3390/educsci13040410

Long D, Magerko B (2020) What is AI Literacy? Competencies and Design Considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems:1–16. https://doi.org/10.1145/3313831.3376727

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654):187–192. https://doi.org/10.1126/science.adh2586

Pennycook G, Rand DG (2021) The Psychology of Fake News. Trends in Cognitive Sciences 25(5):388–402. https://doi.org/10.1016/j.tics.2021.02.007

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems:1–52. https://doi.org/10.1145/3411764.3445315

Rohrer JM, Hünermund P, Arslan RC, Elson M (2022) That's a Lot to Process! Pitfalls of Popular Path Models. Advances in Methods and Practices in Psychological Science 5(2):25152459221095827. https://doi.org/10.1177/25152459221095827

Rollwage M, Loosen A, Hauser TU, Moran R, Dolan RJ, Fleming SM (2020) Confidence drives a neural confirmation bias. Nature Communications 11(1):2634. https://doi.org/10.1038/s41467-020-16278-6

Rollwage M, Fleming SM (2021) Confirmation bias is adaptive when coupled with efficient metacognition. Philosophical Transactions of the Royal Society B 376(1822):20200131. https://doi.org/10.1098/rstb.2020.0131

Rönkkö M, Cho E (2022) An Updated Guideline for Assessing Discriminant Validity. Organizational Research Methods 25(1):6–47. https://doi.org/10.1177/1094428120968614

Schemmer M, Kuehl N, Benz C, Bartos A, Satzger G (2023) Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces:410–422. https://doi.org/10.1145/3581641.3584066

Schulz L, Rollwage M, Dolan RJ, Fleming SM (2020) Dogmatism manifests in lowered information search under uncertainty. Proceedings of the National Academy of Sciences 117(49):31527–31534. https://doi.org/10.1073/pnas.2009641117

Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, Cheng N, Durmus E, Hatfield-Dodds Z, Johnston SR, Kravec S, Maxwell T, McCandlish S, Ndousse K, Rausch O, Schiefer N, Yan D, Zhang M, Perez E (2024) Towards Understanding Sycophancy in Language Models. Proceedings of the International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies 146:102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Spitale G, Biller-Andorno N, Germani F (2023) AI model GPT-3 (dis)informs us better than humans. Science Advances 9(26):eadh1850. https://doi.org/10.1126/sciadv.adh1850

Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour 8(12):2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, Li X, Jin Y, Gašević D (2024) Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology 55(1):90–112. https://doi.org/10.1111/bjet.13370

Zhai C, Wibowo S, Li LD (2024) The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learning Environments 11(1):28. https://doi.org/10.1186/s40561-024-00316-7

Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency:295–305. https://doi.org/10.1145/3351095.3372852

Figure legends

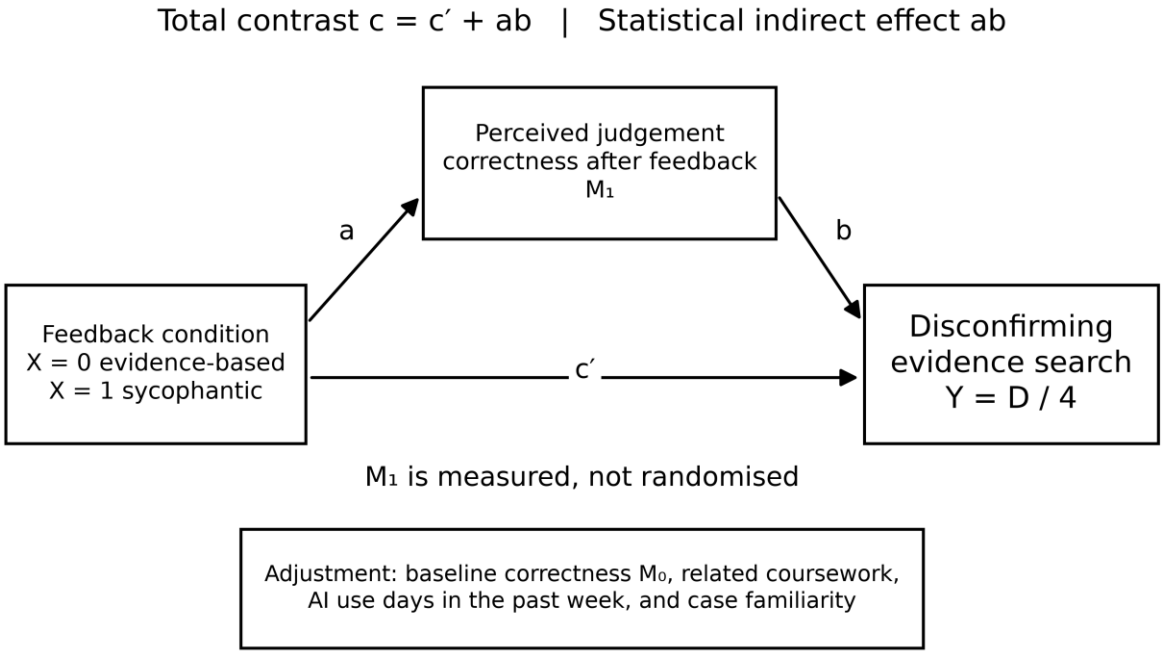


Figure 1. Conceptual and statistical model. X denotes assigned feedback; M_0 and M_1 are perceived judgement correctness before and after feedback; Y = D/4 is the opposing share of four openings. The decomposition uses a common complete-item sample and the same covariates. M_1 is observed rather than randomised. P and Y_s are manipulation-check and auxiliary measures and do not enter the path equations.

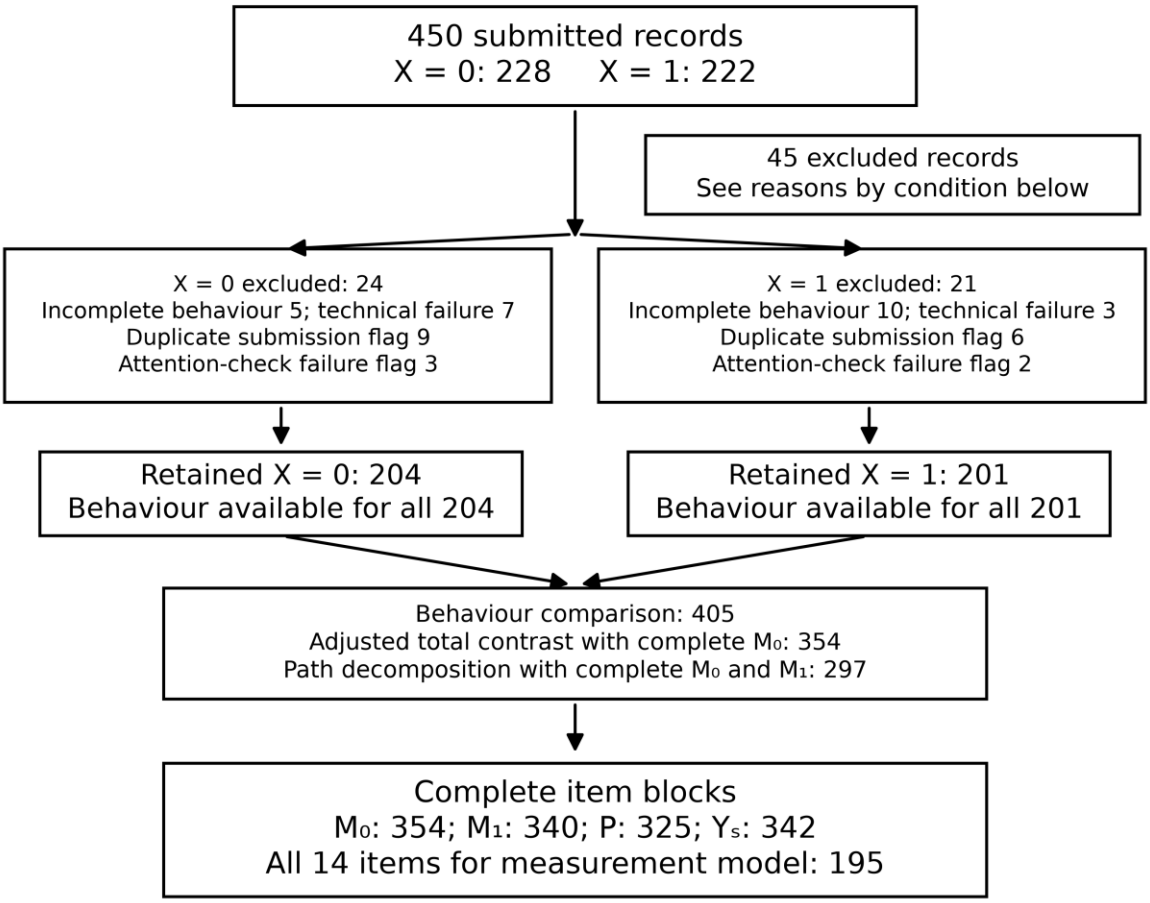


Figure 2. Record flow and analysis samples. Exclusion counts reproduce raw completion codes. Submitted records include observations flagged as duplicates. Complete-item scoring explains the different sample sizes; the 195-person measurement sample is a subset of the retained records.

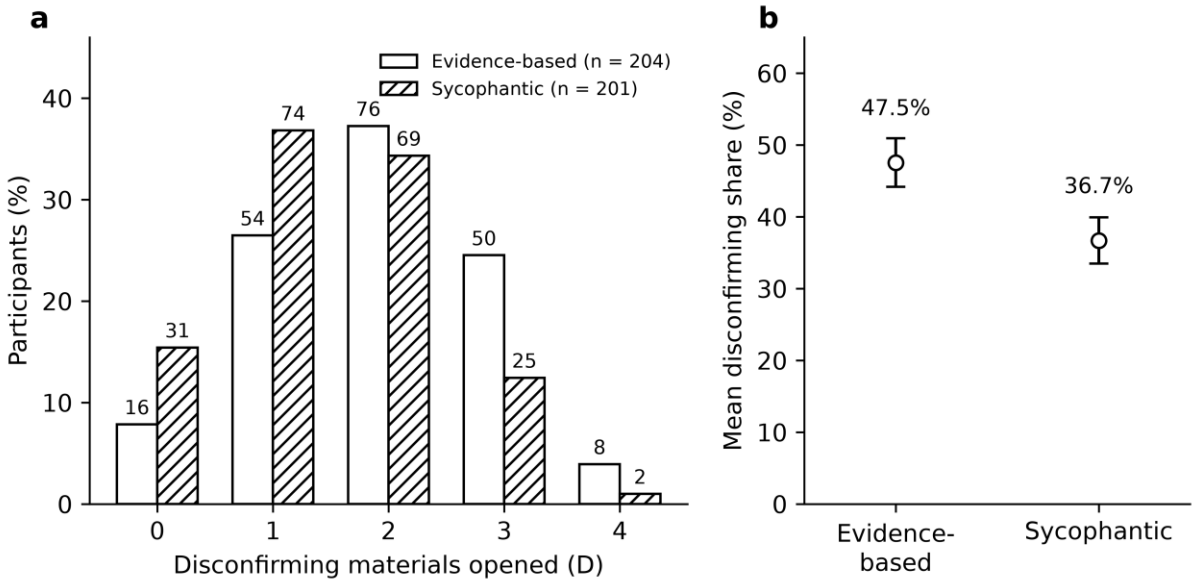


Figure 3. Recorded disconfirming evidence search. Panel a gives the percentages opening zero to four opposing materials; labels show counts. Panel b gives mean Y as a percentage with normal-reference 95% intervals from participant-level means and standard errors. White and hatched bars distinguish conditions. Coding is relative to the initial position; openings do not measure reading depth.

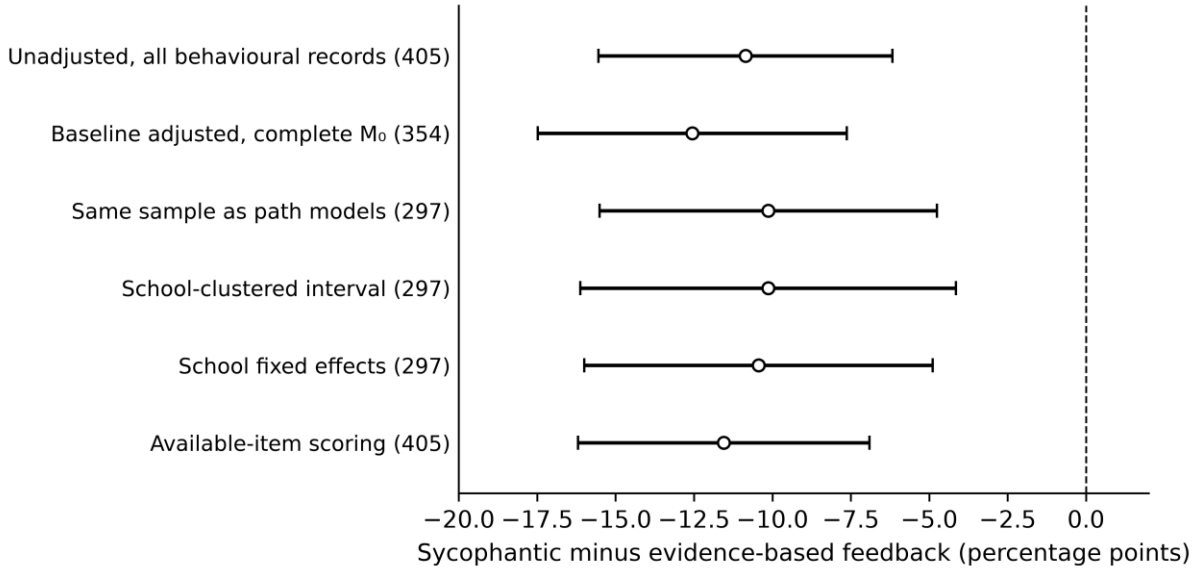


Figure 4. Sensitivity of the condition contrast. Points and 95% intervals are percentage-point differences in Y, sycophantic minus evidence-based feedback. HC3 inference is used except for the school-clustered interval, which uses a finite-sample correction and t(15). Parentheses give sample sizes. Complete-item scoring is used except in the labelled available-item analysis. The analyses overlap and are not independent replications.

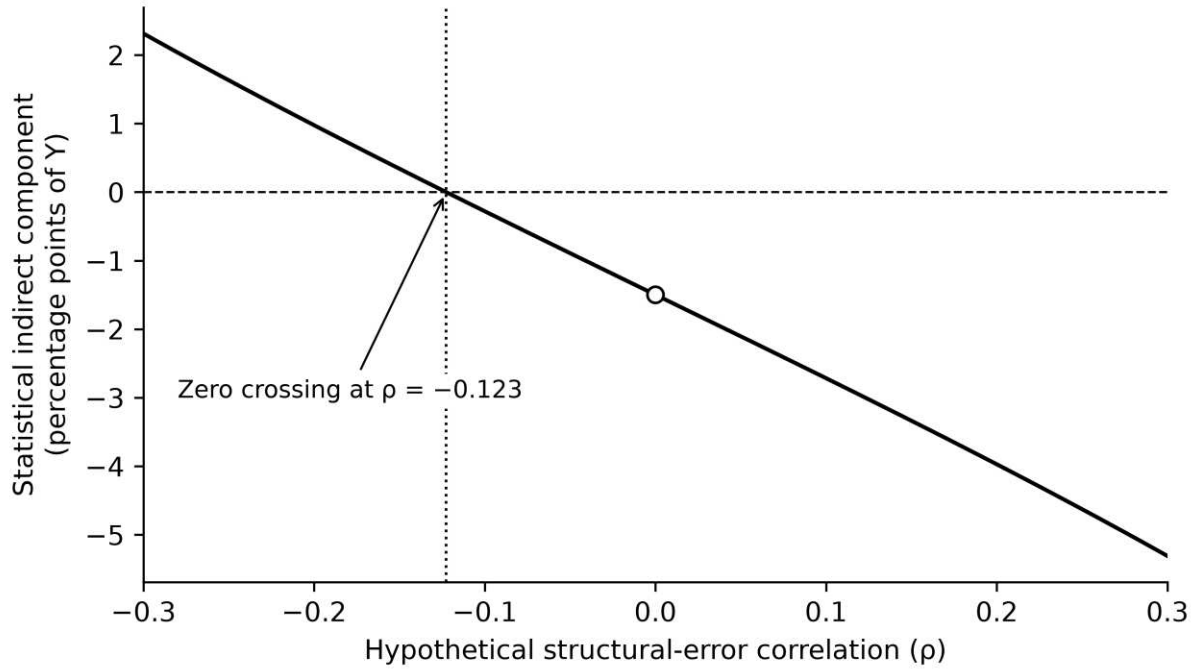


Figure 5. Sensitivity of the statistical indirect point estimate to hypothetical structural-error correlation ρ. The solid curve shows a times b(ρ) under the linear structural-error specification in Supplementary S11. The circle marks ρ = 0; the vertical dotted line marks the zero crossing at approximately −0.123.

Tables

Table 1. Characteristics of the retained sample

Characteristic	Evidence-based n = 204	Sycophantic n = 201	Total n = 405
Henan	53 (26.0%)	54 (26.9%)	107 (26.4%)
Jiangsu	46 (22.5%)	58 (28.9%)	104 (25.7%)
Shandong	53 (26.0%)	48 (23.9%)	101 (24.9%)
Guangdong	52 (25.5%)	41 (20.4%)	93 (23.0%)
Undergraduate year 1	36 (17.6%)	41 (20.4%)	77 (19.0%)
Undergraduate year 2	51 (25.0%)	53 (26.4%)	104 (25.7%)
Undergraduate year 3	58 (28.4%)	50 (24.9%)	108 (26.7%)
Undergraduate year 4	36 (17.6%)	36 (17.9%)	72 (17.8%)
Postgraduate	23 (11.3%)	21 (10.4%)	44 (10.9%)
Relevant coursework completed	115 (56.4%)	119 (59.2%)	234 (57.8%)
Initial position A	112 (54.9%)	107 (53.2%)	219 (54.1%)
AI use days in previous week	3.29 (1.90)	3.51 (1.87)	3.40 (1.88)
Prior case familiarity	3.94 (1.25)	3.85 (1.29)	3.89 (1.27)

Categorical entries are n (% within the column). Continuous entries are mean (SD). Coursework and position rows show one category; complementary categories can be recovered from the denominators. Major counts appear in Supplementary Table S1.

Descriptive statistics and reliability

Table 2. Composite measures and behavioural outcomes

Measure	Evidence-based Mean (SD); n	Sycophantic Mean (SD); n	α	ω	AVE
M_0 Baseline correctness	5.619 (1.329); n = 181	5.437 (1.273); n = 173	0.883	0.893	0.736
M_1 Post-feedback correctness	5.554 (1.594); n = 166	6.079 (1.490); n = 174	0.915	0.919	0.791
P Perceived sycophancy	3.633 (1.093); n = 164	4.823 (1.057); n = 161	0.891	0.893	0.676
Y_s Self-reported search	4.199 (1.036); n = 170	3.997 (1.054); n = 172	0.843	0.834	0.557
D Disconfirming count	1.902 (0.988); n = 204	1.468 (0.933); n = 201	—	—	—
Y Disconfirming share	0.475 (0.247); n = 204	0.367 (0.233); n = 201	—	—	—

Complete-item scoring is used. Alpha sample sizes: M_0 354; M_1 340; P 325; Y_s 342. Standardised omega and average variance extracted (AVE) use the complete-case CFA sample (n = 195), not the full retained sample. Ranges: M_0/M_1 1–9; P/Y_s 1–7; D 0–4; Y 0–1. Internal-consistency coefficients do not apply to X or the behavioural count.

Table 3. Complete distribution of disconfirming material openings

D	Y = D/4	Evidence-based	Sycophantic	Total n
0	0.00	16 (7.8%)	31 (15.4%)	47
1	0.25	54 (26.5%)	74 (36.8%)	128
2	0.50	76 (37.3%)	69 (34.3%)	145
3	0.75	50 (24.5%)	25 (12.4%)	75
4	1.00	8 (3.9%)	2 (1.0%)	10

Every retained participant opened four distinct materials. Percentages use condition-specific denominators of 204 and 201. The coding refers to the initial position, not the final position.

Path models

Table 4. Full path-model coefficients on a common sample

Predictor	Y Total model	M_1 Correctness model	Y Direct/path model
Intercept	0.544 [0.370, 0.717] p < 0.001	1.207 [0.346, 2.067] p = 0.006	0.575 [0.398, 0.751] p < 0.001
X Sycophantic condition	-0.101 [-0.155, -0.048] p < 0.001	0.580 [0.317, 0.842] p < 0.001	-0.086 [-0.142, -0.030] p = 0.003
M_1 Post-feedback correctness	Not included	Not included	-0.026 [-0.049, -0.003] p = 0.027
M_0 Baseline correctness	-0.013 [-0.034, 0.008] p = 0.210	0.818 [0.707, 0.930] p < 0.001	0.008 [-0.020, 0.036] p = 0.588
Relevant coursework	0.035 [-0.019, 0.089] p = 0.207	-0.071 [-0.334, 0.192] p = 0.597	0.033 [-0.021, 0.087] p = 0.230
AI use days	0.012 [-0.002, 0.026] p = 0.103	-0.004 [-0.077, 0.070] p = 0.921	0.012 [-0.002, 0.026] p = 0.105
Case familiarity	-0.018 [-0.039, 0.003] p = 0.094	-0.037 [-0.135, 0.061] p = 0.457	-0.019 [-0.040, 0.002] p = 0.077
n	297	297	297
R^2	0.070	0.507	0.084

Entries are unstandardised B [95% CI] and two-sided p values based on HC3 standard errors and a normal reference distribution. X = 0 is the evidence-based reference. All equations use the same 297 participants and the same pre-feedback covariates. The adjusted total contrast for the larger M_0-complete sample (n = 354) is separately reported in Table 5 and Supplementary Table S6.

Statistical indirect component ab = -0.0150. The 5,000-resample participant-bootstrap percentile 95% interval is -0.0315 to -0.0015. The decomposition is $c = c' + ab = -0.08636 - 0.01498 = -0.10134$.

Sensitivity and auxiliary comparisons

Table 5. Condition contrasts and key inferential sensitivity

Analysis	n	B [95% CI]	p
Unadjusted behavioural contrast	405	-0.1086 [-0.1555, -0.0617]	<0.001
Baseline-adjusted total contrast	354	-0.1256 [-0.1748, -0.0763]	<0.001
Common-sample total contrast	297	-0.1013 [-0.1551, -0.0476]	<0.001
School-clustered total contrast	297	-0.1013 [-0.1612, -0.0415]	0.003
School fixed-effects total contrast	297	-0.1044 [-0.1599, -0.0489]	<0.001
Available-item total contrast	405	-0.1155 [-0.1619, -0.0691]	<0.001
Clustered mediator–outcome association	297	-0.0258 [-0.0533, 0.0016]	0.063
Perceived sycophancy contrast	325	1.1904 [0.9560, 1.4248]	<0.001
Self-reported search contrast	342	-0.2014 [-0.4236, 0.0207]	0.076

Y coefficients are proportion units. P and Y_s coefficients are on their 1–7 scales. School-clustered intervals use 16 schools, a finite-sample correction and t(15); the other linear-model intervals use HC3 with a normal reference. The clustered mediator–outcome row estimates b, not the condition contrast. Available-item scoring differs from the principal complete-item rule.

Ordered-logit sensitivity using all five D levels and the common n = 297 sample

Predictor	Log odds	95% CI	Odds ratio	p
X Sycophantic condition	-0.653	[-1.091, -0.215]	0.520	0.003
M_1 Post-feedback correctness	-0.195	[-0.381, -0.008]	0.823	0.040

The ordered-logit model adjusts for M_0, coursework, AI use days and familiarity. Estimates use the model's conventional maximum-likelihood standard errors and assume proportional odds; they are supplementary, not replacements for the primary mean-difference models.

Additional choice and mechanism analyses

Table 6. Conditional choice contrast and mechanism sensitivity

Analysis	n	λ	Odds multiplier [95% CI]	p
Conditional choice with participant-robust inference	405	-0.7805	0.458 [0.326, 0.645]	<0.001
Conditional choice with school-clustered inference	405	-0.7805	0.458 [0.332, 0.632]	<0.001

Both rows estimate the same 15-parameter conditional likelihood, with separate material preferences for each initial position and m8 as the within-position reference. The model conditions on four selected materials. The odds multiplier compares the relative preference for an opposing versus supporting material between feedback conditions, not the marginal probability of an opening. School inference uses 16 universities and t(15). All coefficients appear in Supplementary S10.

Structural-error sensitivity of the indirect component

Quantity	Value
Common sample	297
Indirect component at $\rho = 0$	-0.01498
Hypothetical ρ at which indirect point estimate is zero	-0.12264

ρ is stipulated rather than estimated. This is a deterministic sensitivity analysis under the stated linear structural-error model. It does not supply a confidence interval, establish the presence of omitted-variable bias or identify a causal mediator.

Data availability

The anonymised dataset, variable codebook and analysis code that support the findings of this study are included as Supplementary Information files accompanying this paper.

Ethics declarations

Ethical approval

The task used a fictional business scenario, collected no sensitive or identifying information, and exposed participants to no more than minimal risk; the simulated AI feedback formed part of the experimental task rather than a concealed manipulation. The study protocol was reported to the research administration office of Zhengzhou Business University under the institution's standard filing procedure for student-participation research involving minimal risk; no separate ethics-committee approval was required under that procedure. The study was conducted in accordance with the Declaration of Helsinki (1975, revised 2013).

Informed consent

Participants provided informed consent, as confirmed by the investigator. Participants were informed that their responses would be anonymised, that participation was voluntary, that no course credit or grade was contingent on taking part, and that they could withdraw at any time without penalty.

Author contributions

Pengfei Li: conceptualisation, investigation, data curation, formal analysis, interpretation, writing of the original draft, and review and editing.

Competing interests

The author declares no financial or non-financial competing interests.

Funding

This research received no specific grant from a public, commercial or not-for-profit funding agency.

Supplementary Files

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

- HSSCSourceDataandAnalysis.zip