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.

AI Assisted Case Analysis and Self Reported Evaluative Confidence among Law Students in Bangladesh

Dr Kazi Abdul Mannan

drkaziabdulmannan@gmail.com

Center for Academic & Professional Career Development and Research (CAPCDR)
https://orcid.org/0000-0002-7123-132X

Md. Hanif Iqbal Hawlader

Shanto-Mariam University of Creative Technology https://orcid.org/0009-0002-1366-5637

Research Article

Keywords: legal education, self-regulated learning, source verification, generative artificial intelligence, evaluative confidence, Bangladesh

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11649675/v1

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

Additional Declarations: The authors declare no competing interests.

AI Assisted Case Analysis and Self Reported Evaluative Confidence among Law Students in Bangladesh

Md. Hanif Iqbal Hawlader

Department of Law, Shanto-Mariam University of Creative Technology, Dhaka, Bangladesh

Email: hanifiqbal7878@gmail.com

ORCID: 0009-0002-1366-5637

Prof. Dr Kazi Abdul Mannan

Adjunct Professor, Department of Business Administration, Shanto-Mariam University of Creative Technology, Dhaka, Bangladesh

Email: drkaziabdulmannan@gmail.com

ORCID: 0000-0002-7123-132X

Corresponding author: Prof. Dr Kazi Abdul Mannan

Abstract

Generative artificial intelligence can support legal case analysis, but perceived assistance, reported verification and confidence in evaluating reasoning represent different educational outcomes. This study examines their alignment in a Bangladesh questionnaire database containing 385 consenting, AI-aware adults. The primary analysis concerns the 237 respondents identifying as law students; the full database provides descriptive context and sensitivity analysis. Low internal consistency supported item-level analysis rather than a composite critical-reasoning scale. Endorsements, paired contrasts, joint response patterns and exploratory regression were used to distinguish reported practices from self-assessed capability. Among students, 66.7% reported checking AI-generated legal information against authoritative sources, 41.4% reported comparing alternative interpretations, 48.9% expressed confidence in detecting weak or unsupported reasoning and 52.3% reported retaining independent judgement. Verification endorsement exceeded detection confidence by 17.7 percentage points (bootstrap 95% interval 9.3-26.2; Holm-adjusted $p < 0.001$). Only 18.6% endorsed verification, detection confidence and independent judgement together. None of four focal predictors-comparison, legal-source verification, awareness of plausible inaccuracies and prompting capability-was retained after multiplicity correction in the adjusted student model. Ordinal, full-sample and secondary judgement-outcome analyses likewise retained no focal coefficient. Self-regulated learning supplies a process-oriented interpretation, while technology acceptance helps delimit perceptions of assistance. The findings identify uneven alignment among self-reports rather than an observed deficit in legal reasoning or a causal effect of AI. Educational implications concern evaluating source checks, argument comparison and independent revision through performance-based tasks, with explicit attention to jurisdiction and authority. The study supports more discriminating assessment of AI-assisted legal learning rather than inferring competence from use or confidence alone.

Keywords: legal education; self-regulated learning; source verification; generative artificial intelligence; evaluative confidence; Bangladesh

1. Introduction

Legal case analysis requires a learner to identify material facts, determine the issues, locate relevant authority and justify the application of rules. An apparently plausible answer is insufficient if it cites the wrong authority or overlooks an alternative interpretation. Scholarship on legal reasoning and professional education treats analytical judgement as more than efficient information retrieval. AI assistance raises a specific educational question within this setting: how do learners assess the reasoning they receive, and how does that assessment relate to the checking practices they report?[1]

Generative systems can help organise information and propose arguments, but their performance must be evaluated at the level of the task and tool. A randomized study of GPT-4 assistance found substantial speed benefits alongside smaller, inconsistent improvements in the quality of legal work. A separate experiment distinguished assistance through highlighting from summarisation, finding different efficiency effects. These results make it inappropriate to infer a general improvement in reasoning from a broad report that AI is helpful.[2]

Recent work also shows that findings change across systems. A 2026 randomized legal-task study of retrieval-augmented and reasoning tools reported stronger quality and productivity results than earlier studies of older models. That evidence strengthens the need to identify which system and task are being evaluated. A questionnaire referring generally to AI cannot be treated as an evaluation of those specific products or as evidence that their benefits transfer to a different educational setting.[3]

The immediate problem is the interpretation of self-report. A student can say that legal information is checked against authoritative sources without expressing the same confidence in detecting unsupported reasoning. Another may report independent judgement while seldom comparing generated arguments with competing interpretations. These answers may reflect distinct practices, differences in question difficulty or uncertainty about capability. Their alignment cannot be assumed from a shared label such as critical legal reasoning.

The supplied questionnaire makes those distinctions visible. It asks separately about issue identification, argument comparison, detecting weak reasoning, systematic examination and independent judgement. A related block asks about legal research and source verification. Rather than collapsing those answers into a single measure, this article examines their distributions and

[1] Frederick Schauer, *Thinking Like a Lawyer: A New Introduction to Legal Reasoning* (Harvard University Press 2009); William M Sullivan and others, *Educating Lawyers: Preparation for the Profession of Law* (Jossey-Bass 2007).

[2] Jonathan H Choi, Amy B Monahan and Daniel Schwarcz, 'Lawyering in the Age of Artificial Intelligence' (2024) 109 Minnesota Law Review 147; Aileen Nielsen and others, 'Building a Better Lawyer: Experimental Evidence That Artificial Intelligence Can Increase Legal Work Efficiency' (2024) 21 Journal of Empirical Legal Studies 979.

[3] Daniel Schwarcz and others, 'AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice' (2026) 3 Journal of Law & Empirical Analysis, doi:10.1177/2755323X261427048.

joint patterns. Research on self-assessed knowledge provides a reason for this caution: self-assessment is not interchangeable with a cognitive performance measure.[4]

The distinction is consequential for educational decisions. A course that records frequent claimed verification may still lack evidence that students can explain why a legal proposition is supported or reject an inapplicable authority. Conversely, neutral confidence responses should not be described as demonstrated failure. The appropriate question for this dataset is whether reported practices and capability judgements align, followed by identifying what direct assessment would be needed to evaluate competence.

Jurisdiction matters because legal authority is structured rather than generic. Article 111 of Bangladesh's Constitution establishes the binding effect of law declared by the Supreme Court's divisions within the judicial hierarchy. This makes the source and status of an authority relevant to evaluating legal propositions. An AI-generated account cannot be assessed solely by its fluency or by confirming that a cited case exists; its relevance and authority also require examination.[5]

The study does not audit students' knowledge of article 111 or of any particular doctrine. It examines self-reports from a legally oriented Bangladesh database. Its population boundary is explicit: 385 adults are recorded, but only 237 identify as current law students. The primary analysis addresses those students, while the remaining records are retained for database descriptions and mixed-population sensitivity analysis. This distinction prevents a broad project database from being mislabelled as an entirely student sample.

Experimental legal-education research evaluates work products under controlled conditions. For example, an examination study of AI assistance found outcome-specific effects rather than an improvement across every component. Research evaluating legal AI tools also documents inaccuracies, including in systems designed to retrieve legal material. These studies justify attention to evaluation and verification but do not establish what students in Bangladesh actually do when using AI.[6]

The present contribution concerns a different evidential level: the relationship between stated routines and evaluative confidence. It neither substitutes for a performance experiment nor claims that legal-AI research is absent. It asks whether practices commonly invoked as safeguards are endorsed by the same respondents, and whether they are associated with the particular capability of detecting unsupported reasoning. Reporting a weak relationship is informative if it prevents a stronger educational claim than the measures warrant.

---

[4] Traci Sitzmann, Katherine Ely, Kenneth G Brown and Kristina N Bauer, 'Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure?' (2010) 9 Academy of Management Learning & Education 169.

[5] Constitution of the People's Republic of Bangladesh, art 111.

[6] Jonathan H Choi and Daniel Schwarcz, 'AI Assistance in Legal Analysis: An Empirical Study' (2025) 73 Journal of Legal Education 384; Matthew Dahl and others, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024) 16 Journal of Legal Analysis 64; Varun Magesh and others, 'Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools' (2025) 22 Journal of Empirical Legal Studies 216.

The general objective is to examine reported case-analysis practices and evaluative confidence among law students in the supplied Bangladesh database. The first question concerns the distribution of assistance perceptions, comparison, detection confidence and independent judgement. The second asks whether legal-source verification, detection confidence and judgement are jointly endorsed. The third examines the associations of comparison, verification, inaccuracy awareness and prompting capability with detection confidence after adjustment for training and generic AI-use frequency. Alternative outcome coding, the full sample and a secondary judgement outcome assess the reach of the main interpretation.

The distinctive analytical contribution is a separation of reported verification from self-assessed evaluation, supported by paired and joint distributions before regression. The approach retains neutral answers, reports unsupported expectations and distinguishes the database population from the student-focused analysis. It yields evidence about alignment among self-reports rather than a claim that AI improves or diminishes legal reasoning. The following sections develop the theoretical distinctions, document methods, present the estimates and consider their educational implications and limitations.

2. Literature Review and Theoretical Foundation

This section links case analysis to self-regulated learning while distinguishing perceived utility from demonstrated competence. Verification, comparison and judgement can occupy different places in a learning process; they need not be interchangeable indicators of one attitude. The conceptual framework therefore follows item wording and uses theory to guide interpretation rather than claiming that the questionnaire fully operationalises established learning models.

2.1 Case analysis as an evaluative activity

Case analysis involves identifying a proposition, evaluating its authority and explaining its application to facts. A student may locate a relevant decision but still misinterpret its reasoning or overlook a distinction. Schauer's account of legal reasoning provides a basis for treating authority and argument as structured forms of analysis. The professional-education literature also connects intellectual work with practical judgement and responsibility. Those perspectives explain why retrieving an answer and evaluating its adequacy should remain distinguishable educational activities.

AI can change the material presented to the learner without resolving those activities. A generated issue list can supply a starting point; a proposed argument can invite comparison. The educational value depends on what the student subsequently does with the output. General educational scholarship on large language models identifies opportunities alongside the need to address accuracy and pedagogical design. Such scholarship supplies a rationale for examining use processes, but not an assumption that access itself produces learning.[7]

[7] Enkelejda Kasneci and others, ‘ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education’ (2023) 103 Learning and Individual Differences 102274.

2.2 Reported practice and self assessed capability

AULR3 asks about checking AI-generated legal information against authoritative sources. ACR2 asks about comparing alternative legal interpretations. Both describe reported practices. ACR3 instead asks whether a respondent can detect weak or unsupported reasoning. This is a capability judgement. ACR5 describes retaining independent judgement when accepting or rejecting generated reasoning. Its endorsement is a report of agency, not a demonstration that the resulting decision is correct.

Self-assessment can capture meaningful perceptions while differing from externally measured knowledge. Sitzmann and colleagues' work cautions against interpreting self-assessed knowledge as direct evidence of cognitive learning. Accordingly, neither high confidence nor low confidence is sufficient to diagnose performance. Without a task and a scoring criterion, this study cannot measure calibration, overconfidence or underconfidence. It can examine whether related statements are jointly endorsed and statistically associated.

The distinctions also affect interpretation of negative or neutral responses. A neutral answer to "I can detect" may indicate uncertainty rather than inability. A respondent might strongly value independent judgement without claiming consistent success in identifying every unsupported argument. Scoring all three as one competence scale would impose an interpretation not established by their content.

2.3 Self regulated learning as the primary lens

Self-regulated learning describes learners as active participants who set goals, apply strategies, monitor performance and reflect on results. Zimmerman's social-cognitive account places these activities in a cycle rather than reducing learning to a stable personal trait. In AI-assisted case analysis, a learner might formulate a question, inspect a proposed answer, check an authority and revise a conclusion. The theory helps identify where evaluation belongs in that sequence.[8]

A reported source check can be interpreted as a monitoring activity, while comparison may support consideration of alternatives. Independent judgement could involve deciding whether a response deserves acceptance or revision. Yet the present questionnaire does not record temporal order, goal setting, feedback use or actual revisions. Its items have conceptual relevance to SRL without constituting a full process measure. The model therefore examines cross-sectional alignment rather than estimating the learning cycle itself.

Zimmerman's overview of self-regulated learners and Panadero's review emphasise the range of cognitive, metacognitive, motivational and contextual elements involved. This breadth matters for interpretation of unretained associations. Weak covariance among a few statements does not refute

[8] Barry J Zimmerman, 'Attaining Self-Regulation: A Social Cognitive Perspective' in Monique Boekaerts, Paul R Pintrich and Moshe Zeidner (eds), *Handbook of Self-Regulation* (Academic Press 2000) 13.

SRL; it identifies the limited explanatory reach of these self-reports. A stronger study would observe strategy use and adaptation during a defined task.[9]

2.4 Technology acceptance as a complementary boundary

Davis's Technology Acceptance Model distinguishes perceived usefulness and perceived ease of use as beliefs relevant to acceptance. The supplied ACR1 and ACR4 statements concern perceived help with issue identification and systematic examination. They have a utility-related meaning, but they are not a complete validated TAM scale. The study contains no full ease-of-use measure or tested acceptance pathway. TAM therefore helps delimit assistance perceptions rather than providing a latent model to be declared confirmed.[10]

This boundary is particularly useful in legal education. A student might find AI helpful for organising a case while remaining uncertain about evaluating its conclusions. Conversely, a capable evaluator may see little benefit in a particular tool. A favourable utility perception and a capability judgement are distinct propositions. The study treats them separately so that an apparent educational benefit is not inferred merely from a statement of helpfulness.

Behavioural theory also distinguishes intentions from realised action. Although this article's principal outcome is capability confidence rather than adoption intention, the broader measurement lesson is applicable: statements about one's position or planned practice do not directly establish what occurs during a consequential task. Self-report must retain its own evidential status.[11]

2.5 Evidence from legal AI and learning studies

Dahl and colleagues evaluated legal hallucinations in general-purpose models using American case-law questions. Magesh and colleagues examined specialist legal research tools and identified remaining reliability problems despite retrieval-based assistance. These findings make source verification a defensible concern. Their task-specific error estimates should not, however, be transferred to unobserved Bangladeshi tools or used to infer the error rate of this student sample.

Human-AI performance studies also resist a single conclusion. Choi, Monahan and Schwarcz distinguished speed from quality, while Nielsen and colleagues showed that highlighting and summaries can affect task completion differently. More recent reasoning and retrieval tools have produced different results in controlled legal tasks. The relevant implication is that system, task and performance criterion must be specified; none is replaced by an undifferentiated claim that AI is useful or harmful.

---

[9] Barry J Zimmerman, 'Becoming a Self-Regulated Learner: An Overview' (2002) 41 Theory Into Practice 64; Ernesto Panadero, 'A Review of Self-Regulated Learning: Six Models and Four Directions for Research' (2017) 8 Frontiers in Psychology 422.

[10] Fred D Davis, 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology' (1989) 13 MIS Quarterly 319.

[11] Icek Ajzen, 'The Theory of Planned Behavior' (1991) 50 Organizational Behavior and Human Decision Processes 179.

Evidence outside law reinforces the need to separate assisted task performance from learning. Bastani and colleagues' mathematics experiment found that the design of AI support mattered for later unaided performance. This does not establish an effect in legal education, but it demonstrates why success while using a tool should not automatically be interpreted as acquired skill. A subsequent legal-learning study should measure both assisted and unaided outcomes.[12]

A survey of knowledge workers by Lee and colleagues considered confidence and self-reported critical-thinking effort during AI use. Its results concern perceptions in a different population rather than objective legal reasoning. It nevertheless illustrates the value of distinguishing confidence in oneself, confidence in a system and the evaluation activity reported. The present study focuses specifically on legal-source checks, alternative interpretations and detection confidence.[13]

2.6 Exploratory expectations

No preregistration or dated analysis plan preceding inspection of the supplied database was available. The following expectations organise exploratory estimates rather than confirmatory hypothesis tests. Each concerns the association between a one-point increase in a focal item and endorsement of ACR3, holding the other focal statements, training and generic use frequency constant. They are not causal propositions supported merely by their theoretical plausibility.

E1 expects reported comparison of alternative legal interpretations to be positively associated with detection confidence. Comparing arguments may expose omissions or weak inferences and therefore has a monitoring-related meaning. Yet reporting comparison does not establish that alternatives are legally relevant or evaluated correctly. A person may compare frequently precisely because confidence is limited. The expected direction is therefore open to empirical assessment.

E2 expects legal-source verification to be positively associated with detection confidence. Consulting authority can reveal an unsupported proposition, but checking that a source exists differs from assessing what it establishes. Some students may follow a verification routine while remaining uncertain about reasoning; others may report confidence without describing that routine. The association must be estimated rather than built into the definition of a competence score.

E3 expects awareness that AI can produce plausible inaccuracies to be positively associated with detection confidence. Awareness could motivate caution and further checking. It might also make a respondent more aware of personal limits and less confident. The item records a claimed understanding of possible error, not skill at diagnosing a particular error. Both the direction and uncertainty of its estimate must therefore be reported.

---

[12] Hamsa Bastani and others, ‘Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics’ (2025) 122 Proceedings of the National Academy of Sciences e2422633122, doi:10.1073/pnas.2422633122.

[13] Hao-Ping (Hank) Lee and others, ‘The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers’ in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (ACM 2025) art 1121, doi:10.1145/3706598.3713778.

E4 expects prompting capability to be positively associated with detection confidence. Formulating effective prompts may help a learner request alternatives or supporting sources. It can also produce more fluent answers without improving evaluation. The supplied item concerns useful legal research support rather than a tested method of exposing errors. A positive association would require interpretation within that limited scope and would not establish a training effect.

2.7 Proposed framework

Figure 1 links the four focal items with detection confidence, adjusting for training and generic AI-use frequency. The framework estimates associations rather than intervention effects or latent learning pathways. Other ACR items describe perceived assistance and independent judgement. Separate paired and joint analyses examine whether the self-reports coincide within students.

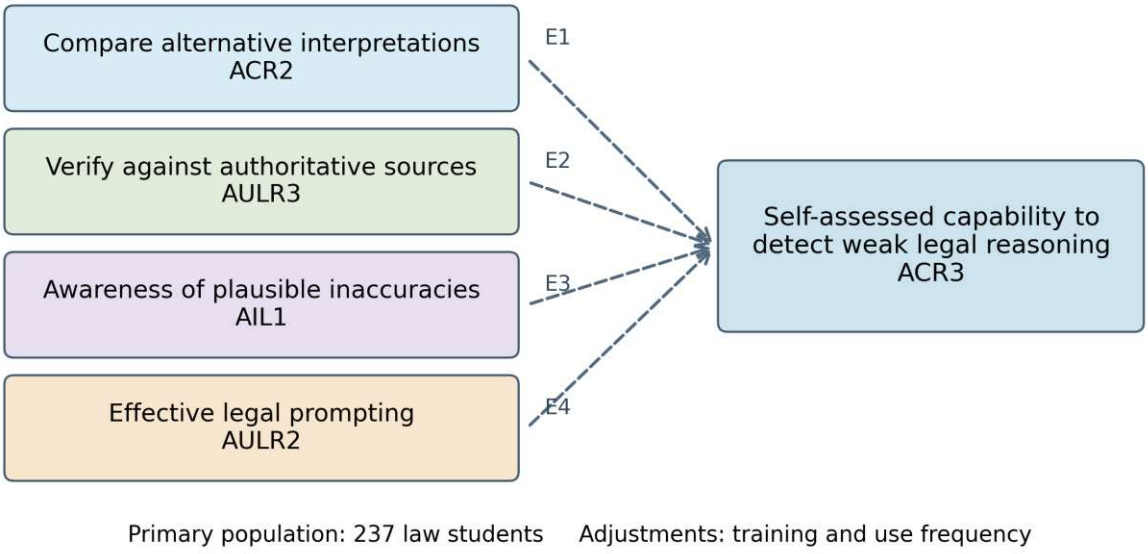


Figure 1. Exploratory framework for reported practices and detection confidence

Dashed arrows represent estimated associations, not causal effects. E1-E4 are exploratory expectations. ACR3 is the primary outcome and ACR5 a separate secondary outcome. No task performance, learning gain or calibration is measured.

3. Methodology

The analysis is anchored in the supplied Group 1 questionnaire and numerical records. The database denominator and the student analytical denominator are kept separate throughout. We use item-level methods because the wording and reliability diagnostics do not justify a homogeneous critical-reasoning score. Collection information, missing documentation and model assumptions are specified below.

3.1 Design and analytical population

This is a cross-sectional questionnaire study in the Bangladesh legal-education context. All 385 records affirm consent, adult eligibility and AI awareness. Current-status responses identify 237 law students, 69 law graduates, 46 advocates or legal professionals and 33 other relevant respondents. The student-focused question is addressed in the 237 current law students. All records remain in full-database descriptions and a mixed-sample model, so the difference in denominators represents a population restriction rather than data-quality exclusions. Reporting follows transparent separation of participant selection, measurement and analysis.[14]

3.2 Sampling and recruitment limits

Recruitment probabilities, institution counts, a sampling frame and invitation totals were not documented. We therefore interpret the achieved sample without national representativeness, design weights or a population response rate. The presence of 385 records does not demonstrate probability sampling or validate a conventional precision calculation. A target of 400, an anticipated non-response rate or an a priori power analysis is not inferred from the article template.

The primary binary model contains four focal predictors and two adjustment variables, with 116 endorsements and 121 other ACR3 responses. The sample supports estimation of this limited specification, but no retrospective claim of adequate power or reliable prediction follows. Participants' institutions are not recorded, preventing cluster-specific uncertainty estimates. Model intervals treat respondent records as independent.

3.3 Questionnaire and item interpretation

The shared instrument covers AI literacy, legal research, case analysis and professional-readiness topics. The same Group 1 database supports related project titles, so this analysis is not an independent replication of companion work. Its distinct focus is alignment between legal-source verification and evaluative confidence. Table 1 identifies the focal outcome and predictors; Appendix A reproduces all 15 analysed or descriptively reported statements. Item-specific validated adaptation sources were not documented, so the theoretical works are not falsely presented as origins of an established scale.

Table 1. Item meanings and analytical roles

Code	Statement focus	Role and interpretation
ACR3	Detect weak or unsupported legal reasoning	Primary outcome; self-assessed capability
ACR2	Compare alternative interpretations	Focal predictor; reported practice
AULR3	Check against authoritative legal sources	Focal predictor; reported practice

[14] Erik von Elm and others, 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies' (2007) 4 PLOS Medicine e296.

Code	Statement focus	Role and interpretation
AIL1	Understand plausible inaccuracies	Focal predictor; claimed awareness
AULR2	Formulate effective legal prompts	Focal predictor; claimed capability
ACR1 and ACR4	Issue identification and systematic examination	Supporting perceptions of assistance
ACR5	Retain independent judgement	Joint patterns and secondary outcome

Source: supplied Group 1 Codebook; exact wording appears in Appendix A. Study-specific statements are not claimed as validated adaptations of complete TAM or SRL scales. No item measures demonstrated reasoning performance.

Responses use categories 1 strongly disagree to 5 strongly agree. ACR3 endorsement means agreement or strong agreement; responses 1-3 form the comparison category, including neutrality. The four focal predictors retain their original 1-5 ordering, assuming a linear log-odds contribution per point. Training combines orientation and course/certified training versus no training. Weekly-or-more use combines weekly and daily/almost-daily generic AI use versus all less frequent categories. Neither measures verified legal-task exposure. ACR5 endorsement is a separate secondary outcome.

3.4 Collection and instrument development

The research team confirmed that the questionnaires were administered in person. Exact fieldwork dates, institutional sites, recruitment channels and invitation counts were not supplied. They are not inferred from the semester or file date. The numerical file also does not identify a specific AI system, subscription or case-analysis task used by each respondent. A completed pilot, translation-validation exercise or cognitive pretest was not documented; no pilot sample or pilot reliability result is claimed.

3.5 Screening and provenance

Blank values and coding ranges were checked while preserving literal categories, including the valid tool response “None”. The 385 records were complete and the analysed agreement items lay within 1-5. No duplicate profile/numerical response was identified after removing row sequence and contextual text. Thirteen records, including seven students, were constant across ACR1-ACR5; none was constant across all 20 agreement items. We retained them because a constant five-item block alone does not establish inattentive responding. Timing and validated attention-check data were unavailable.

The research team identified the numerical records as real questionnaire data and clarified that two Codebook uses of “synthetic” were clerical labels. The contextual-text field was excluded from all substantive analysis, and the sequence field only distinguished rows. No numerical responses were generated, imputed or changed to improve significance or reliability. Data provenance remains dependent on the research records held by the team rather than an independent audit of fieldwork.

3.6 Measurement and statistical analysis

Raw Cronbach alpha was calculated as a scoring diagnostic. In students, values were 0.330 for AI literacy, 0.278 for legal research and 0.198 for case analysis. Together with heterogeneous wording, these results did not justify reflective composites. We therefore report observed items, without invented loadings, composite reliability, AVE, HTMT or latent structural paths. Alpha is not a complete dimensionality test, and item deletion was not used to manufacture a score.[15]

Endorsements use pointwise Wilson 95% intervals. Four student paired contrasts compare AULR3, ACR2, ACR4 and ACR5 with ACR3, using exact McNemar tests and 10,000 respondent bootstrap resamples for percentile difference intervals. Six selected Spearman correlations use 3,000 bootstrap resamples. Separate Holm corrections cover each family. Joint verification-detection-judgement patterns are descriptive rather than latent classes.[16]

The primary student binomial model uses a logit link and HC0 sandwich covariance. Variance inflation factors assess collinearity. A cumulative-logit student model retains ACR3's four observed categories, with common slopes assumed and observed-information intervals. A full-database binary model adds student-status adjustment. A secondary student binary model substitutes ACR5 as outcome. Each has a separate four-focal-coefficient Holm family. ACR5 addresses a related but different question, so it is not treated as an interchangeable replication of ACR3.[17]

3.7 Statistical interpretation

Intervals are pointwise, while adjusted p values address the stated testing families. Neither corrects unknown recruitment bias or measurement error. Exploratory control associations are distinguished from the focal expectations; no claim of an effective training intervention is made. An unretained coefficient is not evidence of equivalence or absence of a mechanism. Cross-sectional estimates cannot determine temporal direction, and no out-of-sample prediction or mediation analysis is claimed.[18]

---

[15] Klaas Sijtsma, 'On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha' (2009) 74 Psychometrika 107.

[16] Edwin B Wilson, 'Probable Inference, the Law of Succession, and Statistical Inference' (1927) 22 Journal of the American Statistical Association 209; Quinn McNemar, 'Note on the Sampling Error of the Difference between Correlated Proportions or Percentages' (1947) 12 Psychometrika 153; Bradley Efron and Robert J Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall 1993); Sture Holm, 'A Simple Sequentially Rejective Multiple Test Procedure' (1979) 6 Scandinavian Journal of Statistics 65.

[17] David W Hosmer Jr, Stanley Lemeshow and Rodney X Sturdivant, *Applied Logistic Regression* (3rd edn, John Wiley & Sons 2013); Halbert White, 'A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity' (1980) 48 Econometrica 817; Peter McCullagh, 'Regression Models for Ordinal Data' (1980) 42 Journal of the Royal Statistical Society Series B 109.

[18] Ronald L Wasserstein and Nicole A Lazar, 'The ASA Statement on p-Values: Context, Process, and Purpose' (2016) 70 American Statistician 129; Sander Greenland and others, 'Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations' (2016) 31 European Journal of Epidemiology 337.

3.8 Ethics and participant protection

Ethical approval for this study was obtained from the Center for Academic & Professional Career Development and Research (CAPCDR), an external institutional body, under approval reference CAPCDR: SMUCT/LLB/SS-0020260908-14. External ethical review was sought because Shanto-Mariam University of Creative Technology does not have an institutional ethics committee responsible for reviewing this type of research.

4. Results

The results first establish the analytical populations, then distinguish assistance perceptions, reported practices and capability confidence. Paired and joint analyses concern the 237 students. The full 385-record database supplies context and a mixed-population model, with each denominator stated explicitly. No result represents a marked case-analysis test or a measured learning gain.

4.1 Database and student profile

Table 2 compares current status and training across the full database and student subset. Students constitute 61.6% of the database. The other 148 respondents comprise 69 graduates, 46 advocates or legal professionals and 33 other relevant adults. They remain part of the database; restricting the main analysis to current students is not a claim that their records are invalid.

Table 2. Full database and law student analytical subset

Measure	Category	All N = 385	Students n = 237
Status	Law student	237 (61.6%)	237 (100.0%)
	Law graduate	69 (17.9%)	0 (0.0%)
	Advocate or legal professional	46 (11.9%)	0 (0.0%)
	Other relevant respondent	33 (8.6%)	0 (0.0%)
Training	No	230 (59.7%)	138 (58.2%)
	Yes, short orientation	105 (27.3%)	65 (27.4%)
	Yes, course or certified training	50 (13.0%)	34 (14.3%)

Column percentages use their stated denominators. The 148 non-students remain in database descriptions and mixed-sample sensitivity analysis; they are outside the primary student-focused population, not excluded for poor data quality.

Among students, 93 were aged 21-23 (39.2%), 53 were 24-26 (22.4%), 41 were 18-20 (17.3%), 29 were 31 or older (12.2%) and 21 were 27-30 (8.9%). There were 132 men (55.7%), 99 women (41.8%), five preferring not to say and one selecting other. Dhaka accounted for 102 students (43.0%), district towns for 53, other city corporations for 48, upazila/rural areas for 29 and other

locations for five. The education-level field records 144 undergraduates, 50 graduates, 25 professional-level and 18 postgraduates; current law-student status, rather than undergraduate level alone, defines the analytical subset.

Weekly or daily/almost-daily generic AI use was reported by 172 students (72.6%). No formal training was reported by 138 (58.2%); 65 reported orientation and 34 a course or certified training. Eight students selected never on experience, compared with four on frequency and 12 selecting the valid tool category "None". These questions capture different dimensions and were not treated as interchangeable eligibility screens. All 237 affirm AI awareness, which does not require previous use.

4.2 Case analysis and supporting item distributions

Table 3 reports all five ACR distributions, retaining the original response categories. Student endorsement ranged from 41.4% for comparing alternative interpretations to 52.3% for independent judgement. Detection confidence was endorsed by 116 students (48.9%; Wilson interval 42.6-55.3%). Its remaining responses were 93 neutral and 28 disagree, without strong disagreement. Thus, non-endorsement was predominantly neutrality and must not be labelled demonstrated inability.

Table 3. Case analysis and evaluative statements by analytical population

Item	Student counts 1 / 2 / 3 / 4 / 5	Student mean (SD)	Student endorsed	Student 95% CI %	All endorsed
ACR1	2 / 19 / 96 / 99 / 21	3.50 (0.80)	120 (50.6%)	44.3 to 56.9	181 (47.0%)
ACR2	2 / 16 / 121 / 72 / 26	3.44 (0.81)	98 (41.4%)	35.3 to 47.7	174 (45.2%)
ACR3	0 / 28 / 93 / 99 / 17	3.44 (0.79)	116 (48.9%)	42.6 to 55.3	187 (48.6%)
ACR4	1 / 12 / 102 / 98 / 24	3.56 (0.76)	122 (51.5%)	45.1 to 57.8	200 (51.9%)
ACR5	0 / 16 / 97 / 96 / 28	3.57 (0.79)	124 (52.3%)	46.0 to 58.6	194 (50.4%)

Student n = 237; full N = 385. Endorsement = response 4 or 5. Categories run from strongly disagree to strongly agree. Wilson intervals are pointwise; means and SDs are descriptive. These are self-reports, not reasoning-test scores.

Issue-identification support was endorsed by 120 students (50.6%) and encouragement of systematic examination by 122 (51.5%). Independent judgement had 124 endorsements (52.3%; 46.0-58.6%), 97 neutral responses and 16 disagreements. Comparison had 98 endorsements (41.4%; 35.3-47.7%), with 121 neutral responses and 18 disagreements or strong disagreements. These distributions do not establish a ranked sequence of cognitive difficulty; the items ask about different propositions.

Legal-source verification was endorsed by 158 students (66.7%; 60.4-72.4%), while effective prompting was endorsed by 110 (46.4%; 40.2-52.8%). Inaccuracy awareness had 101 endorsements (42.6%; 36.5-49.0%). Other shared AI items recorded 176 endorsements for generic

verification (74.3%), 168 for bias/incompleteness awareness (70.9%), 164 for confidentiality safeguards (69.2%) and 167 for human oversight (70.5%). They remain distinct self-reports or preferences rather than a verified literacy score.

The full sample's ACR endorsements were 47.0% for issue identification, 45.2% for comparison, 48.6% for detection confidence, 51.9% for systematic examination and 50.4% for judgement. Those percentages describe all 385 adults and are not substituted for the student estimates.

4.3 Paired contrasts

Table 4 tests four within-student contrasts against detection-confidence endorsement. Legal-source verification exceeded detection confidence by 17.7 percentage points (bootstrap interval 9.3-26.2; Holm-adjusted $p < 0.001$). Eighty students endorsed both, 78 verification alone, 36 detection confidence alone and 43 neither. The contrast therefore concerns unequal endorsement within the same respondents, not a comparison of unrelated groups. Among the 158 verification endorsers, 80 (50.6%) also endorsed detection confidence, compared with 36 of the 79 others (45.6%). These conditional percentages use different denominators from the paired difference and are descriptive. No additional contrast test was applied to them; they provide context for the association estimates reported below. Of verification endorsers, 44 (27.8%) endorsed all three statements in the joint analysis.

Table 4. Paired endorsement contrasts with detection confidence

Contrast	Other item only / detection only	Difference pp	95% CI pp	Holm p
AULR3 − ACR3	78 / 36	17.7	9.3 to 26.2	<0.001
ACR2 − ACR3	48 / 66	-7.6	-16.5 to 1.3	0.333
ACR4 − ACR3	62 / 56	2.5	-6.3 to 11.8	1.000
ACR5 − ACR3	59 / 51	3.4	-5.5 to 11.8	1.000

Student n = 237; pp = percentage points. Exact McNemar tests corrected across four contrasts. Pointwise percentile intervals use 10,000 respondent bootstrap resamples. AULR3 is reported legal-source verification; ACR3 is detection confidence. No contrast measures actual accuracy or a learning effect.

Comparison was endorsed 7.6 points less often than detection confidence, but its interval included zero and adjusted p was 0.333. Systematic-examination support differed by +2.5 points and independent judgement by +3.4 points; both adjusted p values were 1.000. Those unretained contrasts are not demonstrations of equivalence. Judgement and detection confidence still differed in 110 students: 59 judgement-only and 51 detection-only endorsements.

4.4 Joint alignment

Table 5 describes the joint pattern of legal-source verification, detection confidence and independent judgement. This analysis asks whether the three propositions are endorsed together,

rather than inferring alignment from their separate percentages. Neutrality remains in the non-endorsement category and does not identify a failure of practice.

Table 5. Joint endorsement of verification detection confidence and judgement

Legal-source verification AULR3	Detection confidence ACR3	Independent judgement ACR5	Count (%)
Yes	Yes	Yes	44 (18.6%)
Yes	Yes	No	36 (15.2%)
Yes	No	Yes	43 (18.1%)
Yes	No	No	35 (14.8%)
No	Yes	Yes	21 (8.9%)
No	Yes	No	15 (6.3%)
No	No	Yes	16 (6.8%)
No	No	No	27 (11.4%)

Student n = 237. Yes = agreement/strong agreement; No includes neutrality. Patterns are descriptive configurations, not validated learner types, evidence of overconfidence or observed failures of verification.

All three were endorsed by 44 students (18.6%). Forty-three (18.1%) endorsed verification and judgement without detection confidence; 36 (15.2%) endorsed verification and detection without judgement; and 35 (14.8%) endorsed verification alone. Of the 79 who did not endorse verification, 21 endorsed detection and judgement, 15 detection alone, 16 judgement alone and 27 neither. Consequently, 114 students endorsed verification without simultaneously endorsing both other statements. This is a response-pattern observation, not proof that their verification was ineffective.

4.5 Reliability and ordinal associations

Student alpha estimates were 0.330 for AIL, 0.278 for AULR and 0.198 for ACR. The corresponding full-sample values were 0.302, 0.283 and 0.209. These diagnostics, combined with item heterogeneity, did not justify latent composite scores. Table 6 reports selected ordinal associations with ACR3 rather than a correlation between assumed competence scales.

Table 6. Ordinal associations with self assessed detection capability

Pair	Spearman ρ	95% CI	Raw p	Holm p
ACR2 / ACR3	0.101	-0.021 to 0.219	0.121	0.728
AULR3 / ACR3	0.072	-0.059 to 0.193	0.272	1.000
AIL1 / ACR3	0.064	-0.061 to 0.190	0.330	1.000
AULR2 / ACR3	0.098	-0.022 to 0.219	0.133	0.728

Pair	Spearman ρ	95% CI	Raw p	Holm p
ACR5 / ACR3	0.061	-0.071 to 0.195	0.348	1.000
ACR4 / ACR3	0.038	-0.092 to 0.163	0.556	1.000

Student n = 237. Pointwise percentile intervals use 3,000 respondent bootstrap resamples. Holm correction covers six selected correlations. Comparisons concern ordinal self-reports; no objective calibration measure is available.

The six correlations ranged from 0.038 for systematic-examination support to 0.101 for comparison. All adjusted p values were at least 0.728, and all pointwise bootstrap intervals included zero. Legal-source verification had rho 0.072 (interval −0.059 to 0.193; adjusted p = 1.000). The substantial paired endorsement difference therefore coexisted with a small ordinal association. These analyses answer different questions: one concerns relative frequency and the other the ordering of individual responses.

4.6 Adjusted model and alternative analyses

The student model converged, with variance inflation factors from 1.028 to 1.073. Table 7 reports all coefficients and uncertainty. None of E1-E4 was retained: each focal Holm-adjusted p value was 0.878. Adjusted odds ratios were 1.15 for comparison, 1.22 for legal verification, 0.91 for inaccuracy awareness and 1.23 for prompting. Their intervals included 1 and remained compatible with a range of association magnitudes.

Table 7. Student model of detection confidence endorsement

Predictor	aOR	95% CI	Raw p	Holm p
Comparison ACR2	1.15	0.82 to 1.61	0.411	0.878
Legal verification AULR3	1.22	0.87 to 1.71	0.255	0.878
Inaccuracy awareness AIL1	0.91	0.66 to 1.25	0.563	0.878
Prompting AULR2	1.23	0.88 to 1.71	0.219	0.878
Any formal AI training	0.84	0.49 to 1.44	0.527	-
Weekly or daily AI use	1.91	1.05 to 3.48	0.035	-

n = 237; detection endorsements = 116; other = 121. Binomial model with HC0 pointwise intervals. Focal estimates are per one-point increase. Holm correction covers four focal coefficients. Training/use references: no training and less than weekly use. Control p values are unadjusted; aOR is not a probability ratio.

The training control was imprecise (aOR 0.84; interval 0.49-1.44). Weekly-or-more use had aOR 1.91 (1.05-3.48; unadjusted control p = 0.035). This is a nominal exploratory control association, not one of the corrected focal expectations or evidence that frequent AI use improves capability. The binary outcome includes 116 endorsements and 121 other responses.

Table 8 distinguishes three additional analyses. The ordinal student model retains the four observed ACR3 categories. The mixed-sample model uses all 385 adults with student-status adjustment. The secondary student model addresses independent judgement rather than detection confidence. They are overlapping analyses, not independent replications.

Table 8. Alternative coding population and judgement outcome

Model	Focal predictor	aOR	95% CI	Holm p
Student ordinal ACR3	Comparison ACR2	1.22	0.90 to 1.64	0.796
	Legal verification AULR3	1.15	0.85 to 1.56	1.000
	Inaccuracy awareness AIL1	1.04	0.78 to 1.38	1.000
	Prompting AULR2	1.15	0.85 to 1.55	1.000
Full mixed binary ACR3	Comparison ACR2	1.03	0.80 to 1.34	1.000
	Legal verification AULR3	1.08	0.83 to 1.40	1.000
	Inaccuracy awareness AIL1	1.00	0.78 to 1.27	1.000
	Prompting AULR2	1.12	0.87 to 1.43	1.000
Student binary ACR5	Comparison ACR2	0.76	0.55 to 1.07	0.459
	Legal verification AULR3	1.19	0.85 to 1.67	0.894
	Inaccuracy awareness AIL1	0.90	0.66 to 1.23	1.000
	Prompting AULR2	1.07	0.77 to 1.48	1.000

Student ordinal n = 237; observed ACR3 counts: disagree 28, neutral 93, agree 99, strongly agree 17; common slopes assumed, observed-information intervals. Mixed binary N = 385, endorsements = 187, adds student-status adjustment; HC0 intervals. Student judgement binary n = 237, endorsements = 124; HC0 intervals. Separate four-coefficient Holm families. ACR5 is a different outcome, not an interchangeable replication of ACR3.

No focal coefficient was retained in any additional analysis. Ordinal focal adjusted p values were 0.796-1.000; mixed-sample values were all 1.000; and judgement-outcome values were 0.459-1.000. The mixed-sample weekly-use interval included the null (aOR 1.44; 0.92-2.25). Overall, the robust descriptive result is the higher endorsement of legal-source verification relative to detection confidence; the proposed explanatory associations remain unresolved.

5. Discussion

The main finding is uneven alignment among reported verification, evaluative confidence and independent judgement. Two thirds of students endorsed legal-source verification, but fewer than half endorsed the capability to detect weak reasoning, and fewer than one fifth endorsed

verification, detection and judgement together. These patterns describe self-reports. They do not demonstrate either effective verification or poor legal reasoning, and the adjusted models do not establish the proposed explanatory relationships.

5.1 Verification and detection confidence

The paired 17.7-point contrast provides evidence that the two propositions are not equally endorsed. Source checking concerns an activity a student says is performed; detection confidence concerns a judgement of capability. In an SRL interpretation, monitoring activity and belief about evaluative success may occupy different positions in learning. The contrast is therefore compatible with treating them separately, without claiming that one causally produces the other.[19]

Legal-AI reliability studies make the distinction practically relevant. An existing citation may still fail to support an argument, and a generated legal proposition can be unreliable even when presented in a research-oriented system. Those studies explain why a source check should extend to relevance and support. They do not establish that the 78 verification-only students failed to do so: no output or checking record was examined here.[20]

Several interpretations remain possible. Verification may increase awareness of ambiguity; participants may understand "detect" as a demanding claim; or the items may elicit different response styles. None can be selected as the observed explanation. Direct tasks recording the source consulted, the proposition checked and the reason for acceptance or rejection would be needed to distinguish them. Self-assessment research supports keeping the current conclusion at the level of reported practice and confidence.[21]

5.2 Alternative interpretation and independent judgement

Only 41.4% endorsed comparing AI-generated arguments with alternative interpretations. Most other responses were neutral, so the result should not be described as refusal to consider alternatives. It identifies uncertainty or limited endorsement of this specific practice. The comparison statement also leaves frequency, task difficulty and quality of the alternatives unspecified. A course-level diagnosis of critical thinking would require more detailed evidence.

Independent judgement was more frequently endorsed, but 110 students differed between judgement and detection-confidence endorsement. These are conceptually different claims: a person can retain authority over a conclusion while being uncertain about identifying weak reasoning. Legal-reasoning scholarship provides a basis for evaluating the reasons supporting that conclusion rather than treating ownership of the decision as evidence that the reasoning is adequate.[22]

[19] Zimmerman, Attaining Self-Regulation (n 8); Panadero (n 9).

[20] Dahl and others (n 6); Magesh and others (n 6).

[21] Sitzmann and others (n 4).

[22] Schauer (n 1); Sullivan and others (n 1).

The result also qualifies an easy reading of human oversight. A learner's involvement does not by itself establish effective evaluation. Comparative professional guidance, such as ABA Formal Opinion 512, discusses independent review and responsibility for generated legal work. It is not binding law for Bangladeshi students, but it illustrates why meaningful evaluation requires more than simply being the final person who accepts an output.[23]

5.3 Perceived assistance and demonstrated learning

Approximately half of the students endorsed issue-identification assistance and systematic examination. These perceptions should not be discounted, but they are different from measured improvement. TAM helps interpret perceived utility as a belief about the tool's contribution. It does not convert agreement that AI helps into a demonstrated educational effect, particularly when no baseline or comparison condition exists.[24]

Experimental legal research shows why the outcome needs specification. Earlier GPT-4 work found clearer speed benefits than consistent quality benefits, and experiments with highlighting and summaries differentiated forms of assistance. Later reasoning and retrieval tools have generated stronger results on specified legal tasks. The present questionnaire cannot adjudicate among those findings because it records neither the system nor a common assessed task. Its contribution is to clarify the self-reported evaluative positions that a future task-based study should examine.[25]

The distinction between assisted production and learning extends beyond law. Mathematics evidence shows that AI support design can affect later unaided outcomes. That result motivates a legal-education experiment with both assisted and unaided assessment; it does not justify claiming learning loss in these respondents. The current study has no temporal or performance evidence for either benefit or harm.[26]

5.4 Unsupported explanatory expectations

None of the four focal associations was retained in the student binary model, and alternative coding or populations did not produce retained support. These results limit an explanatory account that would infer detection confidence from verification, comparison, awareness and prompting. They do not demonstrate that the activities are educationally irrelevant. The measures are brief, heterogeneous and self-reported, and the intervals leave uncertainty about their relationships.

The findings also do not refute self-regulated learning. SRL concerns a broader process involving goals, monitoring, feedback and adaptation. The study observed no sequence of activity and no change in strategy after an error. A stronger test would examine whether students revise

[23] American Bar Association Standing Committee on Ethics and Professional Responsibility, *Generative Artificial Intelligence Tools* (Formal Opinion 512, 29 July 2024).

[24] Davis (n 10).

[25] Choi, Monahan and Schwarcz (n 2); Nielsen and others (n 2); Schwarcz and others (n 3).

[26] Bastani and others (n 12).

conclusions when source checks reveal unsupported propositions. Without that process evidence, a label such as “SRL confirmed” would exceed the data.

Nominal control findings require similar restraint. Weekly-or-more use was associated with binary detection confidence in the student model, but its mixed-sample interval included the null. Frequency may capture experience, task selection or unrecorded learner differences. An observational control estimate cannot establish that more AI use improves evaluative skill. Statistical guidance supports examining uncertainty and design rather than selecting a favourable threshold result.[27]

5.5 Joint patterns and the population boundary

The 18.6% joint endorsement figure identifies a narrower configuration than any single marginal percentage. It should not be used to label the other students irresponsible: non-endorsement includes neutrality, and a single questionnaire cannot identify stable learner types. Its value is descriptive. It shows how much a broad claim about coordinated verification, confidence and judgement would depend on assuming that different responses represent one construct.

The population boundary is equally important. A 385-record project database is not automatically a 385-student study. Restricting the primary analysis to 237 recorded students makes the educational interpretation clearer, while retaining all records for context avoids concealing the wider sample. Graduates and practitioners may interpret case analysis through different experiences; treating them as interchangeable with current students would obscure that distinction.

5.6 Measurement and integrated interpretation

Common-method influences remain possible because predictors and outcomes were collected through the same questionnaire. Low alpha does not prove their absence, and weak correlations do not authenticate every response. Research on common-method bias supports future improvements such as behavioural corroboration, differentiated measurement and separation of assessment occasions where appropriate.[28]

A survey of knowledge workers has likewise distinguished confidence and reported evaluation effort during AI use. The present results add a specific legal-learning distinction, but neither study converts self-report into direct evidence of critical-thinking performance. The integrated interpretation is therefore bounded: the students' reports distinguish verification from detection confidence and judgement, while a stable explanatory relation among the selected items was not established. This directs further research towards observable checking and revision rather than a larger collection of unvalidated attitude scores.[29]

---

[27] Wasserstein and Lazar (n 18); Douglas G Altman and J Martin Bland, ‘Absence of Evidence Is Not Evidence of Absence’ (1995) 311 BMJ 485.

[28] Philip M Podsakoff and others, ‘Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies’ (2003) 88 Journal of Applied Psychology 879.

[29] Lee and others (n 13).

6. Theoretical and Practical Implications

The implications concern how AI-assisted legal learning should be conceptualised and assessed. They follow from the observed distinctions among statements and the lack of retained explanatory associations. Proposed classroom practices are evaluation priorities, not interventions demonstrated effective by this survey.

6.1 Theoretical implications

First, the study distinguishes reported monitoring activity from evaluative confidence. Treating legal-source verification and detection confidence as separate propositions reveals a paired contrast that a composite would hide. An SRL account can accommodate the distinction, but a future process model must observe what students check and how they respond to the result. The contribution is refinement of measurement, rather than validation of a complete SRL cycle.

Second, independent judgement should be separated from correctness. The student may control the final choice without demonstrating that the chosen argument is well supported. A more discriminating model would connect agency, evidence evaluation and revision through direct measures. The existing questionnaire supports their descriptive separation, while the low reliability and unretained coefficients limit stronger claims about a common competence construct.

Third, AI literacy should not be inferred from exposure or a single self-report score. Long and Magerko's competency-oriented account gives a reason to examine understanding and interaction in context. For legal learning, future measurement should specify an authority, a proposition and a verification decision. Generic prompting confidence and awareness of possible error are relevant but incomplete components.[30]

6.2 Educational and assessment implications

Educators can use the verification-confidence contrast to design tasks that distinguish finding a source from evaluating its support. A student could be asked to locate the cited authority, identify the passage supporting the proposition, establish its jurisdiction and explain a material qualification. The resulting evidence would show which stage needs instruction. The present survey has not demonstrated that students fail any of those steps.

A second task could require comparison of two plausible interpretations followed by a justified acceptance or rejection of an AI-generated argument. Assessment criteria could consider relevance of authority, treatment of facts and explanation of alternatives. Requiring a short revision account would make the student's reasoning visible without assuming that a polished final answer reveals how it was produced.

These assessments should retain separate measures of confidence and performance. A learner could report confidence before and after checking an authority, then complete an unaided task

[30] Duri Long and Brian Magerko, 'What Is AI Literacy? Competencies and Design Considerations' in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM 2020) 1.

later. The design would permit analysis of calibration and transfer, neither of which can be established here. It would also identify whether perceived usefulness corresponds to improved work or only to a smoother drafting experience.

Instruction should explain tool-specific limitations and allow supervised practice with generated errors. Comparative ABA guidance underscores professional responsibility for review, while the legal-AI reliability literature gives concrete reasons to inspect support. Such material can inform teaching without being presented as binding professional rules for every student or as evidence that a particular classroom intervention is already effective.[31]

6.3 Institutional and policy implications

The study offers no national estimate of reasoning competence and no basis for ranking institutions. Curriculum decisions should combine student reports with assessed work, documented recruitment and instructor judgement. In the Bangladesh context, authority-checking exercises should recognise the constitutional structure of precedent rather than treating all retrieved cases as interchangeable. Article 111 provides a concrete doctrinal anchor for such exercises, although students' knowledge of it was not tested here.[32]

Institutions could evaluate AI integration using both assisted work and later independent performance, while documenting the tool, task and support available. Shared project data should also be disclosed when related publications address different questions. These practices would make educational claims more reviewable and prevent a perception survey from being used as proof of improved professional competence.

7. Conclusion Limitations and Future Research

This article examined AI-assisted case analysis through the reported practices and evaluative confidence of 237 law students within a 385-adult Bangladesh database. Two thirds endorsed checking AI-generated legal information against authoritative sources, compared with 48.9% endorsing detection capability and 52.3% reporting independent judgement. Verification exceeded detection confidence by 17.7 percentage points, and only 44 students endorsed verification, detection and judgement together.

These distributional findings coexisted with small ordinal associations and no retained focal regression coefficient after correction. Alternative coding, a mixed-sample model and a secondary judgement outcome did not establish the proposed explanatory relationships. The appropriate conclusion is that reported safeguards and evaluative confidence are unevenly aligned in this sample. The study does not show that AI causes stronger or weaker legal reasoning, that verification is ineffective or that neutral confidence responses demonstrate inability.

The contribution is a transparent distinction between a project database and a student population, and between claimed practice, perceived assistance and self-assessed capability. Paired and joint

[31] ABA, Formal Opinion 512 (n 23); Magesh and others (n 6).

[32] Constitution of Bangladesh, art 111.

analyses recover information obscured by a single score. Preserving those distinctions allows self-regulated learning and technology acceptance to inform interpretation without claiming fully measured latent constructs.

The practical value is to identify what a stronger assessment should observe: the authority checked, the proposition evaluated, the alternative considered and the reason for revision. Those observations would move the evidence from general self-report towards the quality of legal work and independent learning.

Recruitment probabilities, institutional coverage, exact fieldwork dates and invitation totals were unavailable. The achieved student sample is not nationally representative. Generic awareness and use do not verify case-analysis experience or identify the AI system used. Single-item self-reports cannot measure objective reasoning, calibration or learning gains. Reliability diagnostics discourage composite scoring. Cross-sectional data cannot establish direction or causal mechanisms; unknown institutional clustering is unmodelled. Ordinal common slopes were assumed rather than independently verified. Real-data provenance follows the research team's clarification and is not an independent fieldwork audit.

The broader 385-record model supplies a population-sensitivity check, not additional student observations or an independent replication. Complete numerical responses remove the need for imputation but do not eliminate selection bias, common-method influences or uncertainty about interpretation.

Future studies should document sites, dates, recruitment and tool-specific exposure before making broader educational claims. Cognitive interviews could test how students understand verification, weak reasoning and independent judgement. A task study could then compare claimed practice with observed source checking, legal support and revision. Confidence recorded alongside performance would permit calibration analysis.

A randomized classroom design could compare unguided assistance, structured verification prompts and a no-AI condition, using common legal materials and blinded assessment. Delayed unaided work would test learning rather than only assisted production. Longitudinal observation could examine whether students adapt their checking after feedback. These designs would test the processes left unresolved by the present associations while preserving the distinction between perceived help and demonstrated legal reasoning. Documenting access to authoritative materials would also allow researchers to examine learning conditions instead of assuming that every participant has equivalent resources.

Declarations

Ethical approval and oversight

Ethical approval for this study was obtained from the Center for Academic & Professional Career Development and Research (CAPCDR), an external institutional body, under approval reference CAPCDR: SMUCT/LLB/SS-0020260908-14. External ethical review was sought because Shanto-

Mariam University of Creative Technology does not have an institutional ethics committee responsible for reviewing this type of research.

Informed consent

All analysed records affirm consent and adult eligibility. The questionnaire describes voluntary, anonymous participation and aggregate academic reporting.

Data availability

The research team holds the numerical data and analytical specifications. De-identified material may be requested from the corresponding author, subject to confidentiality and applicable requirements. Contextual text was excluded.

Funding and competing interests

No external funding source or competing interest was disclosed in the study materials supplied for manuscript preparation.

Author contributions

Md. Hanif Iqbal Hawlader: lead student researcher; conceptualisation; preparation of study materials; data organisation and curation; manuscript development. The project was conducted under the supervision of Prof. Dr Kazi Abdul Mannan.

Kazi Abdul Mannan: conceptual and methodological guidance; supervision; interpretation of statistical findings; critical review and editing; corresponding authorship.

Generative AI disclosure

Generative AI assisted literature retrieval, drafting, language editing, organisation and analytical-code development. Statistical estimates were computed programmatically from the numerical survey data. Responsibility for analysis, interpretation and reporting rests with the authors.

Bibliography

Ajzen I, 'The Theory of Planned Behavior' (1991) 50 Organizational Behavior and Human Decision Processes 179

Altman DG and Bland JM, 'Absence of Evidence Is Not Evidence of Absence' (1995) 311 BMJ 485

American Bar Association Standing Committee on Ethics and Professional Responsibility, *Generative Artificial Intelligence Tools* (Formal Opinion 512, 29 July 2024)

Bastani H and others, 'Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics' (2025) 122 Proceedings of the National Academy of Sciences e2422633122, doi:10.1073/pnas.2422633122

Choi JH and Schwarcz D, 'AI Assistance in Legal Analysis: An Empirical Study' (2025) 73 Journal of Legal Education 384

Choi JH, Monahan AB and Schwarcz D, 'Lawyering in the Age of Artificial Intelligence' (2024) 109 Minnesota Law Review 147

Dahl M and others, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024) 16 Journal of Legal Analysis 64

Davis FD, ‘Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology’ (1989) 13 MIS Quarterly 319
Efron B and Tibshirani RJ, *An Introduction to the Bootstrap* (Chapman & Hall 1993)
Greenland S and others, ‘Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations’ (2016) 31 European Journal of Epidemiology 337
Holm S, ‘A Simple Sequentially Rejective Multiple Test Procedure’ (1979) 6 Scandinavian Journal of Statistics 65
Hosmer DW Jr, Lemeshow S and Sturdivant RX, *Applied Logistic Regression* (3rd edn, John Wiley & Sons 2013)
Kasneci E and others, ‘ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education’ (2023) 103 Learning and Individual Differences 102274
Lee H-P and others, ‘The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers’ in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (ACM 2025) art 1121, doi:10.1145/3706598.3713778
Long D and Magerko B, ‘What Is AI Literacy? Competencies and Design Considerations’ in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM 2020) 1
Magesh V and others, ‘Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools’ (2025) 22 Journal of Empirical Legal Studies 216
McCullagh P, ‘Regression Models for Ordinal Data’ (1980) 42 Journal of the Royal Statistical Society Series B 109
McNemar Q, ‘Note on the Sampling Error of the Difference between Correlated Proportions or Percentages’ (1947) 12 Psychometrika 153
Nielsen A and others, ‘Building a Better Lawyer: Experimental Evidence That Artificial Intelligence Can Increase Legal Work Efficiency’ (2024) 21 Journal of Empirical Legal Studies 979
Panadero E, ‘A Review of Self-Regulated Learning: Six Models and Four Directions for Research’ (2017) 8 Frontiers in Psychology 422
Podsakoff PM and others, ‘Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies’ (2003) 88 Journal of Applied Psychology 879
Schauer F, *Thinking Like a Lawyer: A New Introduction to Legal Reasoning* (Harvard University Press 2009)
Schwarcz D and others, ‘AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice’ (2026) 3 Journal of Law & Empirical Analysis, doi:10.1177/2755323X261427048
Sijtsma K, ‘On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha’ (2009) 74 Psychometrika 107
Sitzmann T and others, ‘Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure?’ (2010) 9 Academy of Management Learning & Education 169
Sullivan WM and others, *Educating Lawyers: Preparation for the Profession of Law* (Jossey-Bass 2007)
von Elm E and others, ‘The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies’ (2007) 4 PLOS Medicine e296
Wasserstein RL and Lazar NA, ‘The ASA Statement on p-Values: Context, Process, and Purpose’ (2016) 70 American Statistician 129
White H, ‘A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity’ (1980) 48 Econometrica 817
Wilson EB, ‘Probable Inference, the Law of Succession, and Statistical Inference’ (1927) 22 Journal of the American Statistical Association 209
Zimmerman BJ, ‘Attaining Self-Regulation: A Social Cognitive Perspective’ in Monique Boekaerts, Paul R Pintrich and Moshe Zeidner (eds), *Handbook of Self-Regulation* (Academic Press 2000) 13
Zimmerman BJ, ‘Becoming a Self-Regulated Learner: An Overview’ (2002) 41 Theory Into Practice 64

Appendix A Questionnaire wording

Table A1 reproduces every statement used in the focal and supporting analyses. Their analytical roles follow the wording rather than a claim of validated scale equivalence.

Table A1. Exact wording of analysed statements

Code	Questionnaire statement
ACR1	AI assistance helps me identify the key legal issues in a case.
ACR2	I compare AI-generated arguments with alternative legal interpretations.
ACR3	I can detect weak or unsupported reasoning in AI-generated legal analysis.
ACR4	AI-supported case analysis encourages me to examine facts and legal rules more systematically.
ACR5	I retain independent judgement when accepting or rejecting AI-generated legal reasoning.
AULR1	I can identify appropriate generative AI tools for legal research tasks.
AULR2	I can formulate effective prompts to obtain useful legal research support from AI.
AULR3	I verify AI-generated legal information against authoritative legal sources.
AULR4	Using AI helps me locate relevant legal issues, authorities, or arguments more efficiently.
AULR5	I feel competent in using AI without replacing my own legal judgement.
AIL1	I understand that generative AI can produce plausible but inaccurate information.
AIL2	I verify important AI-generated information before relying on it.
AIL3	I understand that AI outputs may reflect bias or incomplete data.
AIL4	I avoid sharing confidential or sensitive information with AI tools unless appropriate safeguards are in place.
AIL5	I believe meaningful human oversight is necessary when AI affects important legal or rights-related decisions.

Source: supplied Group 1 Codebook. Responses range from 1 strongly disagree to 5 strongly agree. No substantive item is reverse-coded.

ACR3 is the primary outcome. ACR2, AULR3, AIL1 and AULR2 are focal predictors. ACR5 is a secondary judgement outcome; the remaining statements provide descriptive context. These items do not constitute a marked legal-reasoning task or objective knowledge test.

Table of Legislation

Constitution of the People's Republic of Bangladesh, art 111