**SHIHAB JOURNAL OF HUMANITIES**
Volume (2), Issue (3) (Sept, 2026)
ISSN: 3136-2958

**مجلة الشهاب للعلوم الإنسانية**
**المجلد (2) العدد 3 سبتمبر 2026**
ردمد: 3136-2958

# *Human or Machine?*
# *EFL Students' Trust Calibration in AI-Generated English Content: A Study at the Faculty of Languages, Surman*

***ASIM OMAR SALIM ALTARABULSI***
*Faculty of Languages- Surman / University of Sabratha*
asim.altrablisi@sabu.edu.ly

***Received: 10/9/2026 Revised: 13/9/2026 Accepted: 30/9/2026 Publication: 27/10/2026***

## Abstract

Generative artificial intelligence (GenAI) is increasingly used by university EFL students for language learning and academic tasks, yet fluent output may be inaccurate. This study examined self-reported trust and verification practices and students' judgments of selected AI-generated English content among undergraduates at the Faculty of Languages, Surman, University of Sabratha. A quantitative descriptive-correlational design was applied to questionnaire responses from 38 students. The instrument included background questions, ten five-point Likert items, five evaluation tasks with confidence and verification-intention responses, and an optional open-ended item. Descriptive statistics, Cronbach's alpha, and Spearman's rank correlations were used. Mean self-reported trust was 3.36/5 and verification was 3.74/5. Students correctly judged 62.1% of task responses under the primary answer key. Use frequency was not significantly associated with trust ($\rho = .279$, $p = .090$), whereas verification was associated with accuracy ($\rho = .473$, $p = .003$), as was confidence ($\rho = .648$, $p < .001$). Findings are preliminary given the small sample, modest trust-scale internal consistency, and ambiguity in one evaluation item. In exploratory HC3 regression, confidence remained associated with accuracy ($\beta = .51$, $p = .005$), while reported verification did not uniquely reach $p < .05$ after confidence was included ($p = .063$). After Holm correction across 28 exploratory pairs, only confidence–accuracy and awareness–adaptive-trust correlations remained significant. The study highlights the value of teaching verification and interpreting trust relative to output reliability.

*Keywords:* EFL, generative artificial intelligence, trust calibration, verification, AI literacy.

# Chapter One: Introduction

## 1.1 Background of the Study

The rapid development of Artificial Intelligence (AI), particularly Generative Artificial Intelligence (GenAI), has introduced significant changes to the ways students access, process, and produce information. Tools based on large language models can generate human-like English texts, provide explanations, translate texts, suggest vocabulary, correct grammatical errors, summarize information, and assist learners in producing academic and non-academic writing.

The increasing use of GenAI in higher education has created opportunities as well as challenges for students and teachers. In English as a Foreign Language (EFL) contexts, AI-generated content can provide learners with immediate linguistic support and opportunities for independent learning. At the same time, AI-generated responses may contain inaccurate information, inappropriate language, fabricated references, grammatical or contextual problems, or culturally unsuitable expressions. Recent research in L2 writing has therefore highlighted the importance of verification and calibrated trust rather than uncritical acceptance of AI feedback. (Li et al., 2026) (Miao & Holmes, 2023; Ji et al., 2023). A recent EAP intervention study also examined pedagogically integrated GenAI and student AI literacy and engagement (Gill et al., 2026).

Trust calibration is especially important because a fluent response is not necessarily an accurate response. Students may be influenced by the apparent confidence, detail, or naturalness of AI-generated English. A learner who accepts incorrect content with high confidence demonstrates a different pattern from a learner who recognizes uncertainty, checks evidence, and adjusts reliance according to the task. (Lee & See, 2004; Li et al., 2026).

Recent empirical work distinguishes general attitudes toward AI from demonstrated ability to evaluate AI output. A performance-based GenAI literacy assessment found that task performance can provide information beyond self-reported proficiency (Jin et al., 2025). Research in EFL contexts also treats GenAI evaluation as a component of broader AI literacy (Li & Ding, 2025; Wang, Derakhshan, et al., 2025).

For EFL learners, this issue has particular significance. Students may use AI for grammar, vocabulary, translation, writing, speaking preparation, and academic assignments. In each of these situations, the consequences of an inaccurate response may differ. A minor vocabulary suggestion may be easily corrected, whereas an incorrect academic explanation or fabricated reference may have more serious consequences. Effective use therefore requires students to adjust their reliance to the task.

Recent research in university EFL writing has examined trust calibration and verification. A study of EFL writing classes linked classroom norms and verification-oriented practices with how students engaged with AI feedback and revised their writing (Li et al., 2026). A qualitative study of Moroccan graduate students described a cycle of inspecting, diagnosing, refining, verifying, adopting, editing, or rejecting AI feedback (Mhamdi, 2026).

Despite the growing international literature, there is a need for context-specific evidence from Libyan higher education. The present study therefore focuses on undergraduate EFL

students at the Faculty of Languages, Surman, University of Sabratha and examines both self-reported trust and practical responses to AI-generated English content.

## 1.2 Statement of the Problem

Generative AI is increasingly accessible to university students, including EFL students. Students can obtain explanations, translations, corrections, essays, summaries, vocabulary suggestions, and other English-language content almost instantly. However, accessibility and frequent use do not necessarily mean that students can evaluate AI-generated content accurately.

The central problem addressed by this study is the possibility of a mismatch between students' trust in AI-generated English content and the actual reliability of that content. Students may trust a fluent answer too much, fail to verify information, or alternatively distrust useful AI assistance even when it is accurate. Such mismatches are relevant to responsible AI-supported language learning.

At the Faculty of Languages, Surman, there is limited empirical evidence concerning students' levels of trust in AI-generated English content, their verification practices, and their ability to distinguish accurate from inaccurate AI-generated language. The present study addresses this problem by combining a short questionnaire with five practical AI-content evaluation tasks. This design allows the research to examine students' stated attitudes as well as their responses to concrete examples.

## 1.3 Purpose of the Study

This study investigated EFL students' trust calibration in relation to AI-generated English content at the Faculty of Languages, Surman, University of Sabratha.

The study specifically aimed to:

1. measure students' self-reported trust in AI-generated English content;
2. examine students' verification practices;
3. examine students' perceived ability to recognize inaccuracies in AI-generated English;
4. examine students' practical judgments of accurate and inaccurate AI-generated content;
5. examine the relationship between AI use, verification practices, and trust calibration; and
6. provide evidence that may inform responsible use of GenAI in EFL learning.

## 1.4 Research Questions

1. What is the level of EFL students' trust in AI-generated English content?
2. To what extent do EFL students verify AI-generated English content before using it?
3. How accurately do EFL students judge the correctness of selected AI-generated English examples?
4. How confident are students when judging the correctness of AI-generated English content?
5. Is there a relationship between students' AI use/experience and their self-reported trust?
6. Is there a relationship between verification practices and practical trust-calibration performance?

## 1.5 Research Hypotheses

H1: There is a statistically significant relationship between frequency of AI use and self-reported trust in AI-generated English content.

H2: There is a statistically significant relationship between verification practices and practical trust-calibration performance.

H3: Students' confidence in judging AI-generated content is significantly related to the accuracy of their judgments.

The achieved sample was 38. Given the ordinal measures and small sample, the planned associations were examined using two-tailed Spearman rank correlations; findings are interpreted cautiously.

## 1.6 Significance of the Study

For EFL students, the study may increase awareness of the need to evaluate AI-generated content rather than accept it automatically.

For EFL instructors, the findings may provide information useful for classroom activities involving AI, verification, critical thinking, and responsible digital literacy.

For the Faculty of Languages, Surman, the findings may provide preliminary local evidence for discussions about appropriate and responsible integration of GenAI into English-language education.

For researchers, the study may provide a context-specific starting point for future research on AI literacy, trust, verification, critical evaluation, and English language learning in Libya.

## 1.7 Scope of the Study

The study focuses on undergraduate EFL students at the Faculty of Languages, Surman, University of Sabratha. It examines students' trust calibration in relation to English-language content generated or modified by generative AI tools such as ChatGPT and comparable systems.

The study does not evaluate the technical performance of AI systems as a whole. Instead, it examines students' perceptions, verification practices, confidence, and judgments of selected AI-generated English content.

## 1.8 Operational Definitions

Artificial Intelligence (AI): Computer-based technologies capable of performing tasks associated with aspects of human intelligence, including language processing and generation.

Generative Artificial Intelligence (GenAI): AI systems capable of generating new content, including text, explanations, translations, summaries, and other language output.

AI-Generated English Content: English-language text or linguistic information produced, corrected, translated, summarized, or substantially modified by a generative AI system.

Trust in AI: The degree to which a student believes that AI-generated information is reliable, accurate, and suitable for use.

Trust Calibration: The extent to which a learner's confidence and reliance-related judgments correspond to the actual reliability of a particular AI output. In this study, it is examined cautiously through accuracy judgments and confidence across five keyed examples, alongside self-reported verification and item-level verification intention. Because the

questionnaire did not ask whether students would actually adopt each response, adoption/reliance itself was not directly measured.

Verification: The process of checking AI-generated information against reliable sources, teachers, textbooks, dictionaries, or other evidence before accepting or using it.

EFL Students: Undergraduate students studying English as a Foreign Language at the Faculty of Languages, Surman.

# Chapter Two: Literature Review

## 2.1 Introduction

This chapter reviews literature related to AI in education, GenAI in EFL learning, AI-generated English content, trust in AI, trust calibration, verification, AI literacy, and empirical approaches to evaluating students' interaction with GenAI. It concludes by identifying the research gap addressed by the present study.

## 2.2 Artificial Intelligence and Generative AI in Education

AI has increasingly become part of higher education through systems that support information access, personalization, automated feedback, and language learning. GenAI has expanded these possibilities by allowing users to interact with systems through natural language and receive newly generated responses. (Miao & Holmes, 2023; Gill et al., 2026).

The educational significance of GenAI is not limited to efficiency. It also changes the learner's role from a recipient of information to an evaluator and editor of machine-generated content. This shift makes critical evaluation and verification increasingly important. (Long & Magerko, 2020; Miao & Holmes, 2023).

## 2.3 GenAI in EFL Learning

In EFL settings, GenAI can support writing, vocabulary, grammar, translation, speaking preparation, and feedback. Research on Chinese university EFL learners describes GenAI literacy as a process involving learners' uses and evaluation of AI during language tasks (Li & Ding, 2025). A scale-development study in the EFL context identifies knowledge, use, evaluation, design, and ethics as dimensions of teacher GenAI literacy; because its participants were teachers, it informs the construct but does not directly measure student trust calibration (Wang, Derakhshan, et al., 2025). Related intervention research has examined AI-generated multi-tiered EFL reading tasks and reading attitudes/outcomes, though not trust calibration (Alghamdi & Alghizzi, 2026).

Other recent EFL research has examined AI-mediated feedback and learner trust. In a qualitative study of 62 Moroccan graduate students learning English, learners described inspecting outputs, diagnosing weaknesses, refining prompts, and selectively adopting, editing, verifying, or rejecting AI suggestions (Mhamdi, 2026).

These findings support the importance of studying not only whether students use AI but also how critically they evaluate its output.

## 2.4 Accuracy and Limitations of AI-Generated Content

Large language models can produce fluent and plausible responses without guaranteeing that every claim is correct. In language learning, this may appear as grammatical explanations that are incomplete, translations that do not fit the context, vocabulary suggestions that are inappropriate for the intended register, or factual information that is inaccurate. (Ji et al., 2023; Miao & Holmes, 2023).

This creates an important pedagogical distinction between linguistic fluency and informational or contextual reliability. Students need to understand that natural-sounding English is not sufficient evidence of correctness.

## 2.5 Trust in AI

Trust refers to a willingness to rely on a system under conditions of uncertainty. In automation research, trust is important because it shapes whether reliance is appropriate to the system's capabilities and the task context (Lee & See, 2004). In GenAI-supported language learning, trust can be associated with reliance, resistance, perceived risk, and actual use (Wang, Li, et al., 2025).

However, trust should not be interpreted as an inherently positive outcome. Excessive trust may encourage uncritical acceptance, whereas insufficient trust may prevent students from benefiting from useful assistance. The relevant educational goal is therefore appropriately calibrated trust. (Lee & See, 2004; Wang, Li, et al., 2025).

## 2.6 Trust Calibration

Trust calibration concerns the correspondence between a person's trust and the actual capabilities or reliability of an automated system (Lee & See, 2004). In this study, calibrated trust refers to judgments that are appropriately responsive to whether a particular AI-generated answer is correct, with confidence and verification intention providing additional evidence. This is an operational, task-limited use of the concept, not a claim that the questionnaire provides a fully validated calibration measure.

Recent L2 writing research has directly connected trust calibration with verification practices. A 2026 study of university EFL writing classes found that classroom conditions and verification-oriented practices were associated with how learners engaged with AI feedback and with deeper revision. (Li et al., 2026)

This concept is central to the present study because it moves beyond the simple question "Do students trust AI?" and instead asks whether their trust is appropriately related to whether an AI-generated answer is correct.

## 2.7 Verification as a Metacognitive Safeguard

Verification involves checking AI-generated information against another source or evidence before relying on it. It may include consulting a textbook, dictionary, teacher, reliable website, academic source, or another independent reference.

Research in higher education has examined students' verification strategies and reported that verification quality can be limited, supporting the need for explicit checking routines (Martín-Moncunill & Alonso Martínez, 2025). EFL feedback research also describes selective verification as part of learners' interaction with GenAI suggestions (Mhamdi, 2026).

For EFL learners, verification can be particularly important because students may lack sufficient linguistic or subject knowledge to identify subtle errors. Verification therefore functions as a practical bridge between trust and critical use.

## 2.8 AI Literacy and Evaluation Skills

AI literacy involves more than operating an AI tool; it includes understanding AI, critically evaluating its outputs, and interacting with it in informed ways (Long & Magerko, 2020). In EFL contexts, GenAI literacy has been discussed in relation to knowledge, use, evaluation, design, and ethics (Wang, Derakhshan, et al., 2025; Li & Ding, 2025).

A scale-development study with EFL teachers identified GenAI knowledge, use, evaluation, design, and ethics as distinct dimensions (Wang, Derakhshan, et al., 2025). A performance-based assessment study found that direct assessment of GenAI literacy predicted performance on GenAI-supported tasks beyond self-reported proficiency (Jin et al., 2025).

These findings provide methodological support for the present study's decision to combine Likert-scale self-report items with practical evaluation tasks.

## 2.9 Practical Evaluation of AI Output

Self-report questionnaires are useful for measuring perceptions, attitudes, and reported behaviors, but they may not fully capture actual evaluation ability. A student may report that they always verify AI information but still accept an inaccurate answer when presented with a realistic example.

For this reason, the study included five short evaluation tasks. Some examples were accurate and others contained errors. Students judged correctness, indicated confidence, and stated whether they would verify the response before use. This design allows comparison between reported trust and judgments of concrete outputs, while the task score is interpreted as a limited performance indicator.

The inclusion of direct evaluation tasks is consistent with evidence that performance-based measures can reveal competencies not fully captured by self-reports (Jin et al., 2025).

## 2.10 Human or Machine? The Relevance of Source Perception

The title of the present study raises the distinction between human and machine-generated English because students may interpret an answer differently depending on who or what they believe produced it. Research comparing human and AI-generated educational explanations has investigated how students evaluate explanations when source identity is controlled or disclosed.

In the present study, however, the main focus is not on whether students can identify whether a text was written by a human or AI. Instead, the study focuses on whether students evaluate the content itself appropriately. This distinction is important because a student may trust or distrust a response for reasons unrelated to its actual accuracy.

## 2.11 Previous Research

Recent research provides several relevant strands. EFL-focused studies examine GenAI literacy and learner use (Li & Ding, 2025; Wang, Derakhshan, et al., 2025), while broader higher-education assessments demonstrate the value of performance-based measures (Jin et al., 2025). Studies of EFL writing and English learning address trust, verification, and selective uptake of GenAI feedback (Li et al., 2026; Mhamdi, 2026).

Second, recent research on L2 writing has explicitly connected trust calibration with verification and learning-oriented engagement with AI feedback (Li et al., 2026).

Third, research on AI-mediated English learning indicates that learners may adopt, edit, verify, or reject AI suggestions depending on task and feedback conditions (Mhamdi, 2026).

Fourth, performance-based assessment research suggests that demonstrated GenAI literacy may not be represented adequately by self-reported confidence alone (Jin et al., 2025).

Together, these findings support the use of a mixed measurement approach within a primarily quantitative study.

## 2.12 Research Gap

Although research on GenAI and language learning is expanding rapidly, the present study addresses a more specific gap.

First, the studies reviewed here are situated primarily in Chinese, Hong Kong, Moroccan, and Spanish higher-education settings rather than Libyan undergraduate EFL programmes (Li &

Ding, 2025; Li et al., 2026; Martín-Moncunill & Alonso Martínez, 2025; Mhamdi, 2026; Wang, Li, et al., 2025). The present study therefore contributes preliminary, locally situated evidence from Surman; it does not claim that no prior Libyan research exists.

Second, the reviewed literature includes work on trust and verification as well as studies using direct performance assessments, but there remains value in combining these approaches in a single local EFL study (Jin et al., 2025; Martín-Moncunill & Alonso Martínez, 2025).

Third, the distinction between confidence and correctness matters because performance-based GenAI literacy assessments can provide evidence beyond self-ratings (Jin et al., 2025).

The present study addresses these gaps by combining a concise questionnaire with five practical AI-output evaluation tasks among undergraduate EFL students at the Faculty of Languages, Surman.

## 2.13 Conceptual Framework

The framework links AI-use experience with self-reported trust and verification. It treats trust calibration as the fit between a student's judgments and confidence and the known correctness of each example. AI-output awareness, task-adaptive trust, and academic level are contextual indicators. The arrows represent proposed associations consistent with a descriptive-correlational design, not causal effects.

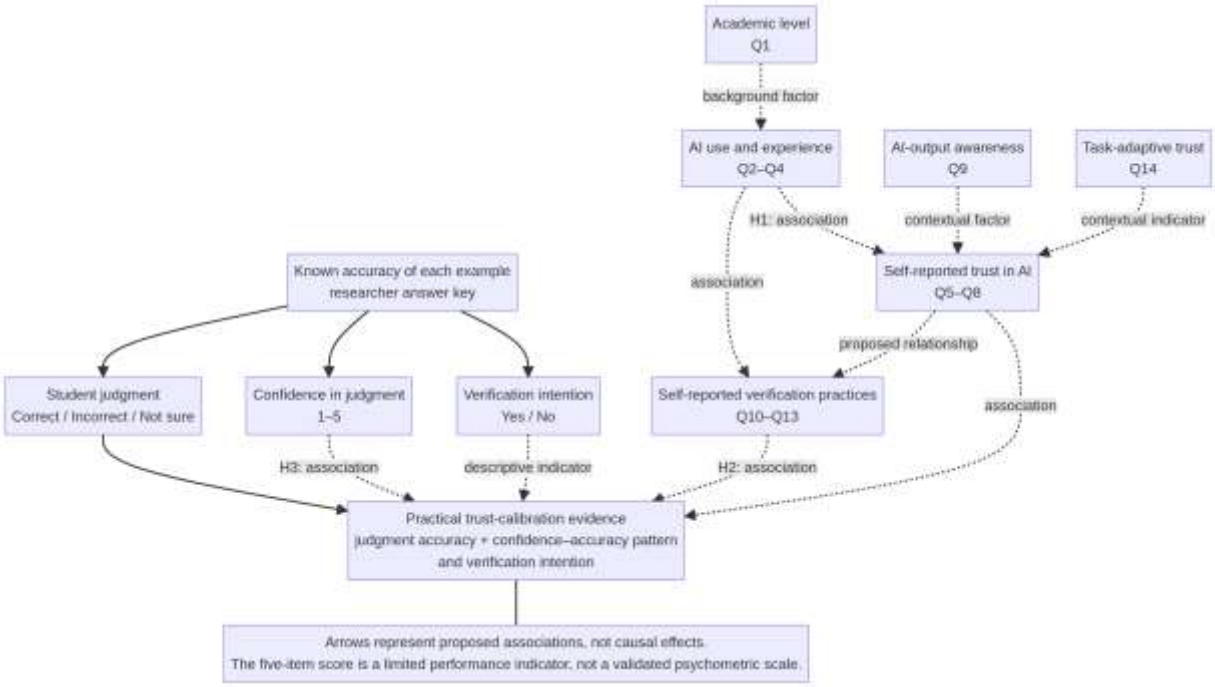



**Figure 1**

*Conceptual Framework for EFL Students' Trust Calibration in AI-Generated English Content*
*Note. Arrows indicate proposed associations rather than causal effects. Practical calibration evidence is limited to five examples and is not a validated psychometric scale; adoption of each answer was not directly measured.*

# Chapter Three: Methodology

## 3.1 Research Design

The study employed a quantitative descriptive-correlational design with a small performance-oriented component. Descriptive analyses summarized self-reported trust, verification practices, and evaluation responses. Spearman rank correlations examined the specified associations. Five practical evaluation items complemented self-report data; the cross-sectional design does not establish causal relationships.

## 3.2 Population and Sample

The target population was undergraduate EFL students enrolled at the Faculty of Languages, Surman, University of Sabratha.

The achieved analytic sample comprised 38 undergraduate EFL students. The questionnaire described participation as voluntary. The supplied files do not document the full sampling frame, recruitment process, response rate, or nonresponse, so representativeness cannot be established. The academic-level distribution is reported in Chapter Four.

## 3.3 Research Instrument

The questionnaire contained 20 numbered questions in three components:

Part A: Background information (Questions 1–4).
Part B: Self-reported AI trust, verification, and critical evaluation (Questions 5–14), using a five-point Likert scale.
Part C: Five practical AI-generated English evaluation tasks (Questions 15–19).
Question 20 will provide a short optional open-ended opportunity for students to state an additional concern or observation.

The tasks were brief to keep the questionnaire manageable. The achieved dataset contained 38 responses.

## 3.4 Practical AI Evaluation Task

Five short examples were presented: a past-time grammar error, a vocabulary/usage error, a paraphrase of an English idiom, an incorrect explanation of the irregular past tense of "go," and an overgeneralized claim about AI grammar. The intended key classified Questions 15, 16, 18, and 19 as incorrect and Question 17 as correct. The wording of Question 17 is only an approximate paraphrase and the resulting sensitivity is reported in Chapter Four.

For each example, participants indicated (a) whether the content was correct, incorrect, or uncertain; (b) confidence in that judgment on a five-point scale; and (c) whether they would verify the answer before using it.

The participant questionnaire did not disclose the answer key.

The answer key was retained for scoring and was not included in the participant questionnaire.

## 3.5 Scoring of Practical Tasks

For each practical item, a judgment matching the researcher's answer key was coded 1; all other judgments, including "Not sure," were coded 0 for the primary accuracy score. "Not sure" responses were also tabulated separately so they were not conflated with explicit incorrect Yes/No beliefs.

Confidence was recorded on a five-point scale. Verification intention was coded 1 for "Yes" and 0 for "No."

The practical accuracy score was the number of keyed judgments divided by five. The score is a limited indicator based on five items, not a validated psychometric measure of trust calibration.

Confidence and accuracy were compared descriptively at the item and participant levels. Because the task contained only five examples, these patterns are interpreted as exploratory indicators rather than as a complete psychometric assessment of trust calibration.

High confidence in an explicitly incorrect judgment was treated as a possible overconfidence pattern. Because the task contains only five items and no item-level adoption decision, these patterns were not used to classify individual students or claim a complete measure of trust calibration.

## 3.6 Questionnaire Response Scale

Questions 5–14 used the following five-point scale:

1 = Strongly Disagree
2 = Disagree
3 = Neither Agree nor Disagree
4 = Agree
5 = Strongly Agree

Questions 5–8 were averaged as a self-reported trust indicator and Questions 10–13 as a verification-practice indicator. Questions 9 and 14 were analyzed separately because they measure awareness and adaptive trust, respectively, rather than the same subscale. No reverse coding was applied to the two composites.

## 3.7 Validity

No documentation of expert review was included in the supplied research files. Accordingly, this study does not claim that content validity was formally established. The ambiguity in the Question 17 paraphrase reinforces the need for specialist review before any replication.

Experts should evaluate relevance, clarity, linguistic appropriateness, difficulty, and alignment with the research questions. The practical examples should also be checked to ensure that the correct answer is defensible and that the intended error is not ambiguous.

Any item that produces substantial disagreement among experts should be revised before the pilot or main study.

## 3.8 Pilot Study

The supplied materials did not document a pilot study; no pilot sample or pilot results are reported.

Because pilot testing was not documented, item clarity and interpretation cannot be assumed. In particular, the Question 17 ambiguity should be resolved in a revised instrument.

Cronbach's alpha was interpreted cautiously as an estimate of internal consistency, not as evidence of construct validity. The observed coefficients are reported in Chapter Four.

## 3.9 Data Collection Procedure

The supplied questionnaire informed students that participation was voluntary, responses were anonymous, and data would be used for academic research. The dataset contained 38 submissions. The exact administration platform, dates, recruitment process, and response rate were not documented in the supplied files.

The questionnaire instructed participants to evaluate each example individually and gave an estimated completion time of 8–12 minutes. The files do not permit confirmation of whether participants completed the tasks independently.

The exported dataset contained timestamps but no names or student numbers; timestamps were excluded from the de-identified analysis file. The files do not identify whether the researcher was also an instructor or how participation was separated from assessment.

## 3.10 Ethical Considerations

The questionnaire stated that participation was voluntary, responses were anonymous, and the data would be used only for academic research. Documentation of formal institutional ethics approval was not supplied; no approval is claimed here.

The questionnaire stated that participation was voluntary and that responses would be used for academic research. The supplied files do not document the data-storage protocol or retention schedule; these should be specified in the final institutional submission.

The principal ethical considerations were informed voluntary participation, anonymity, and responsible handling of research data. Formal ethics-review details were not provided.

### 3.11 Data Analysis

The 38-response CSV was analyzed using Python (pandas and SciPy). Timestamps were excluded from the de-identified analysis copy.

Descriptive analyses summarized background frequencies and percentages, Likert-item and composite distributions, practical-task accuracy, confidence, and verification intention.

Internal consistency was estimated separately for the four-item trust and verification composites using Cronbach's alpha; coefficients are interpreted cautiously.

Two-tailed Spearman rank correlations tested the prespecified associations between use frequency and trust, reported verification and practical accuracy, and task confidence and practical accuracy. The wider exploratory matrix included 28 pairwise correlations, with Holm family-wise adjustment. Exploratory ordinary least squares models estimated the 0–5 participant accuracy score first from verification practice and then from verification plus mean task

confidence; a sensitivity model used the strict alternative key for Question 17. Coefficients are reported with HC3 heteroskedasticity-consistent standard errors and 95% confidence intervals. Because the outcome is bounded/discrete and n = 38, these models are exploratory associations, not causal or externally validated predictions.

No inferential group comparisons were conducted because the academic-year subgroups were small and uneven.

Given the achieved sample of 38 and the five-item performance task, factor analysis, formal psychometric validation, and complex multivariable models were not conducted. The two-predictor OLS analyses reported in Chapter Four are deliberately parsimonious and exploratory; they do not establish causality or validate the five-item score as a measurement scale.

Practical-task data were summarized as accuracy (0–5), mean confidence, and verification intention. Participant-level Spearman correlations were used for ordinal/limited-scale associations; exploratory OLS models estimated the accuracy score from verification alone and from verification plus confidence, with a strict-key sensitivity model for Question 17. These measures distinguish self-reported attitudes and intentions from task performance.

### 3.12 Operational Variable Structure

**Table 1**

*Operationalization of Study Variables*

| Variable | Role in framework | Operational measure |
|---|---|---|
| AI use and experience | Predictor/context | Prior use, frequency, and reported uses (Q2–Q4) |
| Self-reported trust | Main self-report construct | Mean of trust items (Q5–Q8) |
| Verification practice | Main self-report construct | Mean of verification items (Q10–Q13) |
| AI-output awareness | Contextual indicator | Awareness that fluent AI can be wrong (Q9) |
| Adaptive trust | Contextual indicator | Self-reported task-sensitive trust (Q14) |
| Academic level | Background variable | Year of study (Q1) |
| Practical judgment accuracy | Performance indicator | Keyed correctness across five tasks (Q15–Q19) |
| Judgment confidence | Performance indicator | Confidence rating for each task (1–5) |
| Verification intention | Performance indicator | Yes/No response for each task |
| Trust-calibration evidence | Derived, limited outcome | Correspondence of known item accuracy with judgment/confidence patterns; verification intention reported separately |

### 3.13 Study Limitations

The achieved sample of 38 was below the planned range of 50–80, and the academic-year groups were uneven. Voluntary response may have introduced selection bias. Nearly all respondents had some AI experience, limiting variation in the use-frequency measure. The trust composite had modest internal consistency ($\alpha = .63$), while the verification composite was more consistent ($\alpha = .73$); neither coefficient establishes validity. Self-reports may not match actual behavior. Five practical tasks provide only a limited snapshot and Question 17 is ambiguous. One participant's "ever used AI" response conflicted with the reported frequency. Formal validation, pilot, ethics-approval, and sampling-frame documentation were not provided.

The findings therefore provide preliminary local evidence only and should not be generalized beyond the study sample without replication.

## 3.14 Chapter Summary

This chapter described the quantitative descriptive-correlational study, its 20-question instrument, scoring of five performance items, and analysis of 38 responses. Self-reported trust and verification were analyzed separately from practical accuracy, confidence, and verification intention. The interpretation remains cautious because of sample size, scale reliability, item ambiguity, and undocumented validation and sampling procedures.

# CHAPTER FOUR
# DATA ANALYSIS AND RESULTS

This chapter reports the analysis of the 38 questionnaire submissions supplied for analysis. All statistics below are calculated from the supplied CSV. The original instrument and dataset were retained; the analysis plan and reporting were expanded to include exploratory correlation and regression analyses.

## 4.1 Introduction

This chapter presents the findings in relation to the six research questions and three hypotheses. It summarizes respondent characteristics, AI-use patterns, self-reported trust and verification, performance on five evaluation items, associations among the principal variables, and the optional written comments. Because the achieved sample was 38 rather than the proposed 50–80, the results are interpreted as context-specific and exploratory.

## 4.2 Data preparation and scoring

The CSV contained 38 records and no blank cells in the exported responses. Timestamp fields were excluded from the analysis copy. The ten Likert responses were coded from 1 (Strongly Disagree) to 5 (Strongly Agree). Self-reported trust was operationalized as the mean of Questions 5–8. Self-reported verification was the mean of Questions 10–13. Question 9 (awareness that fluent AI output may be wrong) and Question 14 (adjusting trust to task demands) were retained as individual indicators because each represents a distinct single item.

For the primary performance score, the keyed judgments were No for Questions 15, 16, 18, and 19, and Yes for Question 17. A response received one point only when it matched the key. "Not sure" was therefore not counted as a correct judgment, but it is also reported separately from an explicitly incorrect Yes/No judgment. The wording of Question 17 is an approximate paraphrase, so an alternative scoring check is reported in Section 4.5. Spearman's rank correlation was used for the planned ordinal associations. Tests are two-tailed, with $\alpha = .05$.

## 4.3 Respondent profile and use of AI

**Table 2**

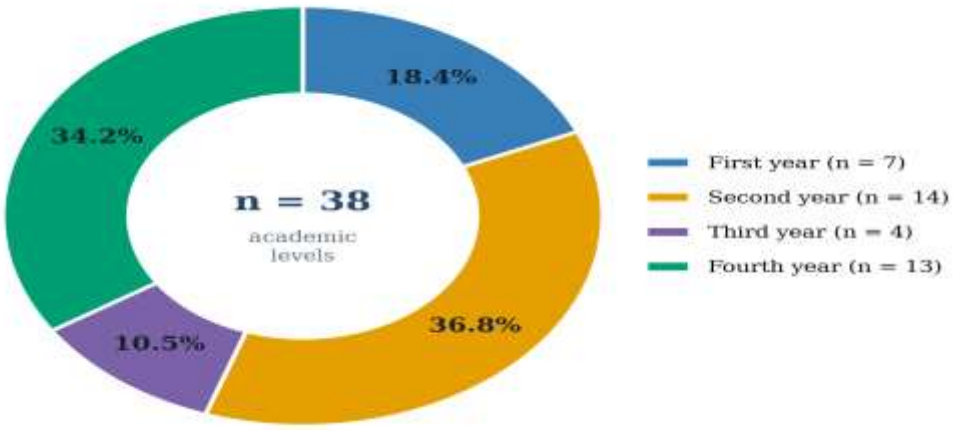
*Respondents by Academic Level*

| Academic level | n | % |
|---|---|---|
| First year | 7 | 18.4 |
| Second year | 14 | 36.8 |
| Third year | 4 | 10.5 |
| Fourth year | 13 | 34.2 |

Most participants were in their second year (36.8%) or fourth year (34.2%). Third-year students were the smallest subgroup (10.5%), so comparisons by academic year would be unstable and are not treated as inferential findings. The respondent distribution is displayed in Figure 2.

**Figure 2**

*Academic-Level Distribution of Respondents*




*Note.* The chart shows counts and percentages for the four academic levels (n = 38); percentages are rounded.

**Table 3**

*Prior Use of AI for English or Academic Work*

| Ever used AI for English/academic work | n | % |
|---|---|---|
| Yes | 37 | 97.4 |
| No | 1 | 2.6 |

AI use was common. Reported frequency was Sometimes for 22 students (57.9%), Often for 6 (15.8%), Very often for 6 (15.8%), and Rarely for 4 (10.5%). One response is internally inconsistent: the respondent answered No to ever having used AI but selected Rarely for frequency. Frequency analyses use the recorded ordinal frequency answer and should be interpreted with this data-quality note.

**Table 4**

*Reported Purposes of AI Use (Multiple-Response Item)*

| Reported use (multiple selections allowed) | n | % of respondents |
|---|---|---|
| Writing | 24 | 63.2 |
| Speaking | 24 | 63.2 |
| Translation | 23 | 60.5 |
| Summarizing | 23 | 60.5 |
| Vocabulary | 14 | 36.8 |
| Grammar | 1 | 2.6 |

Writing, speaking, translation, and summarizing were the most frequently selected uses. Because this was a multiple-response question, percentages do not sum to 100%. Additional free-text uses included writing long paragraphs, generating ideas, and understanding information.

### 4.4 Self-reported trust, awareness, and verification

**Table 5**

*Descriptive Statistics for Self-Report Items*

| Item | Statement (shortened) | M | SD |
|---|---|---|---|
| Q5 | Generally trusts AI-generated English explanations | 3.34 | 0.97 |
| Q6 | Believes AI-generated English answers are usually accurate | 3.29 | 0.69 |
| Q7 | Feels confident using AI content for learning | 3.47 | 1.08 |
| Q8 | Tends to believe clear, confident AI answers | 3.32 | 0.96 |
| Q9 | Knows fluent AI English can still be wrong | 3.92 | 1.02 |
| Q10 | Checks AI information before academic use | 4.05 | 0.90 |
| Q11 | Compares AI answers with reliable sources | 3.53 | 1.27 |
| Q12 | Verifies translations when accuracy matters | 3.76 | 1.08 |
| Q13 | Asks a teacher/reliable source when uncertain | 3.63 | 1.20 |
| Q14 | Adjusts trust to task type and accuracy needs | 3.76 | 1.17 |

The mean self-reported trust score was 3.36 (SD = 0.64) on the five-point scale, indicating moderate-to-positive trust rather than uniformly strong trust. The trust subscale's Cronbach's alpha was $\alpha = 0.63$; this is modest and warrants caution in treating the four-item mean as a highly precise scale.

The mean self-reported verification score was 3.74 (SD = 0.83). Its internal consistency was $\alpha = 0.73$. Students most strongly endorsed checking information before academic use (Q10: M = 4.05); comparison with textbooks, dictionaries, or reliable sources had a lower mean (Q11: M = 3.53). Awareness that fluent English can still be wrong was relatively high (Q9: M = 3.92), as was reported adjustment of trust to task importance (Q14: M = 3.76). These figures represent reported attitudes and practices, not directly observed everyday behavior.

### 4.5 Performance on the five practical evaluation items

**Table 6**

*Performance, Confidence, and Verification Intention by Evaluation Item*

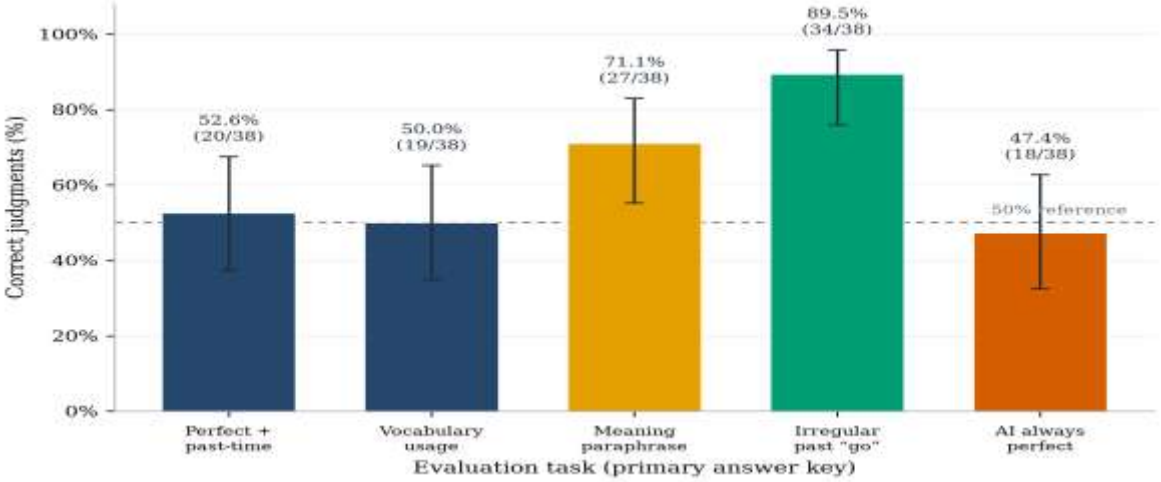
| Q | Evaluation item | Key | Correct | % correct | Mean confidence | Would verify % |
|---|---|---|---|---|---|---|
| 15 | Grammar: present perfect with a finished past-time marker | No | 20/38 | 52.6 | 3.79 | 81.6 |
| 16 | Vocabulary/usage: "interested in" | No | 19/38 | 50.0 | 3.66 | 65.8 |
| 17 | Translation/paraphrase of "looking forward to" | Yes | 27/38 | 71.1 | 3.74 | 57.9 |
| 18 | Irregular past tense of "go" | No | 34/38 | 89.5 | 3.71 | 60.5 |
| 19 | Claim that AI text is always grammatically perfect | No | 18/38 | 47.4 | 3.39 | 68.4 |

Across all 190 item responses, students made 118 keyed judgments correctly (62.1%), giving a mean individual score of 3.11/5 (SD = 1.71; median = 3.5). Performance was strongest on the "goed" item (89.5%) and weakest on the claim that AI is always grammatically perfect (47.4%), followed by the two sentence-correction items (50.0–52.6%). Across item responses, mean confidence was 3.66/5, and participants indicated that they would verify 66.8% of responses.

The task-level verification percentages ranged from 57.9% to 81.6%. The grammar item attracted the most verification intention (81.6%), while the meaning/paraphrase item attracted the least (57.9%). Intent to verify did not itself establish that respondents would conduct an effective check in practice. Figure 3 displays the task-level accuracy estimates with 95% Wilson intervals.

**Figure 3**

*Correct Judgments on the Five Evaluation Items*




*Note.* Bars show the percentage of correct judgments under the primary answer key; error bars are 95% Wilson intervals (n = 38 per item). Question 17 is the paraphrase item whose wording is discussed in the scoring-sensitivity analysis.

Scoring sensitivity for Question 17: the primary key treats the paraphrase as correct because it conveys the central meaning of anticipating a future meeting with positive feeling. If a stricter evaluator instead marks it incorrect because "waiting with excitement" is not an exact or fully natural equivalent, the mean becomes 2.53/5 (50.5%) rather than 3.11/5 (62.1%). Under this alternative key, the verification–score association remains significant ($\rho = .455$, $p = .004$), as does the confidence–score association ($\rho = .533$, $p < .001$). This sizable difference makes the item's intended answer key and expert review important for any final institutional submission.

### 4.6 Hypothesis testing

**Table 7**

*Results of the Three Primary Hypotheses*

| Hypothesis | Spearman result | Decision at α = .05 |
|---|---|---|
| H1: Use frequency is related to self-reported trust | $\rho = .279$, $p = .090$; pHolm(3) = .090 | Not statistically significant |
| H2: Verification practice is related to practical performance | $\rho = .473$, $p = .003$; pHolm(3) = .005 | Supported within the three primary tests; see Table 8 for broader exploratory correction |
| H3: Confidence is related to judgment accuracy | $\rho = .648$, $p < .001$; pHolm(3) < .001 | Supported; also survives the 28-pair Holm correction |

H1 was not supported: the positive association between use frequency and self-reported trust was not statistically significant. H2 showed a positive bivariate association between reported verification and practical accuracy ($\rho = .473$, $p = .003$); this remained significant after Holm adjustment across the three original hypotheses (pHolm = .005), but not in the broader 28-pair exploratory matrix (pHolm = .071). In the multiple regression controlling for task confidence, verification did not make a statistically significant unique contribution ($p = .063$). H3

was supported: task confidence was associated with accuracy and remained a unique predictor in the multiple regression; the confidence–accuracy association also survived the 28-pair Holm correction.

A separate count-based pattern helps illustrate possible overconfidence. Across the 190 item judgments, 28 were "Not sure." Among the 113 judgments with high confidence (4 or 5), 18 were definite incorrect Yes/No judgments (15.9% of high-confidence responses). Uncertain responses are not treated as incorrect beliefs. The 18 high-confidence errors show that strong confidence did not guarantee correctness, particularly on items with lower accuracy. The five-item instrument is too short to classify individual students reliably as over-trusters.

### 4.7 Expanded Correlation Analysis

The full Spearman matrix below summarizes 28 pairwise exploratory associations among eight study indicators. To reduce family-wise error, Holm adjustment was applied across these 28 tests. Only the confidence–accuracy association and the association between risk awareness (Q9) and adaptive trust (Q14) survived this broader adjustment. The original H2 verification–accuracy association remained significant when correcting only the three original hypotheses, but not across the wider exploratory family.

**Table 8**

*Spearman Correlations Among Core Study Variables*

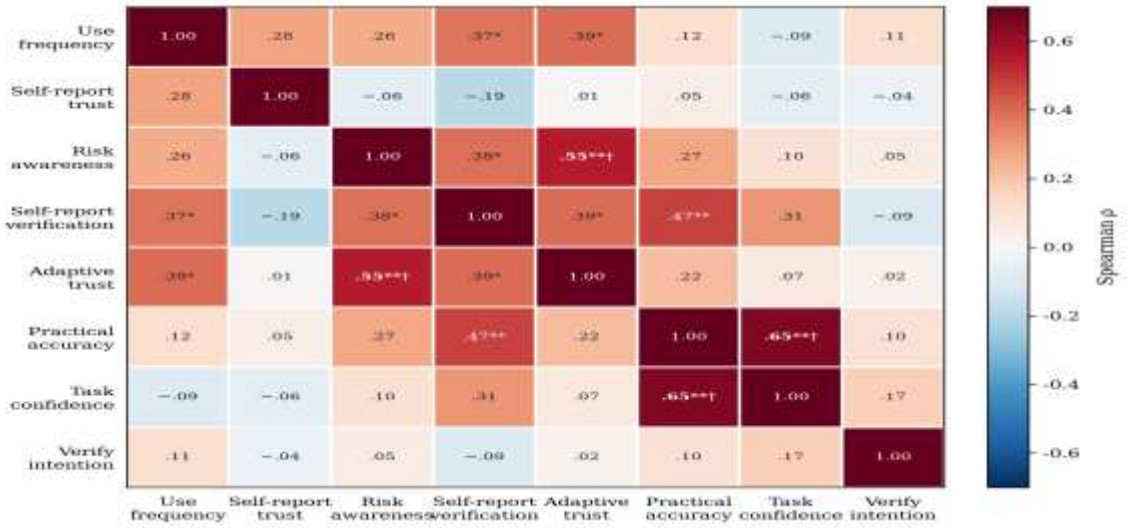
| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 Use frequency | — | .28 | .26 | .37* | .39* | .12 | −.09 | .11 |
| 2 Trust | .28 | — | −.06 | −.19 | .01 | .05 | −.06 | −.04 |
| 3 Awareness | .26 | −.06 | — | .38* | .55**† | .27 | .10 | .05 |
| 4 Verification | .37* | −.19 | .38* | — | .39* | .47** | .31 | −.09 |
| 5 Adaptive trust | .39* | .01 | .55**† | .39* | — | .22 | .07 | .02 |
| 6 Accuracy | .12 | .05 | .27 | .47** | .22 | — | .65**† | .10 |
| 7 Confidence | −.09 | −.06 | .10 | .31 | .07 | .65**† | — | .17 |
| 8 Task verify | .11 | −.04 | .05 | −.09 | .02 | .10 | .17 | — |

*Note.* Entries are Spearman's ρ; two-tailed tests; n = 38. Codes: 1 = AI-use frequency (Rarely = 1, Sometimes = 2, Often = 3, Very often = 4); 2 = mean Q5–Q8 trust; 3 = Q9 risk awareness; 4 = mean Q10–Q13 verification; 5 = Q14 adaptive trust; 6 = primary-key accuracy (0–5); 7 = mean confidence; 8 = proportion of task responses participants said they would verify. *p < .05 and **p < .01 are unadjusted; †p < .05 after Holm adjustment across all 28 pairs.

Confidence was moderately associated with accuracy (ρ = .648, p < .001, Holm p = .0003). Risk awareness was also associated with adaptive trust (ρ = .552, p < .001, Holm p = .0089). Verification and accuracy showed a positive bivariate association (ρ = .473, p = .003), but its Holm-adjusted p across the full matrix was .071. Self-reported trust was not meaningfully associated with practical accuracy (ρ = .049, p = .772), and the association between use frequency and trust remained nonsignificant (ρ = .279, p = .090). Figure 4 visualizes the full Spearman matrix; significance markers follow the correction described in Table 8.

**Figure 4**

*Spearman Correlations Among the Eight Study Indicators*




*Note.* Cell values are Spearman's ρ. *p < .05 and **p < .01 are unadjusted; † identifies pairs surviving Holm adjustment across the 28 pairwise tests.

## 4.8 Exploratory Regression Analysis

Three parsimonious OLS models were estimated with the practical accuracy score (0–5) as the outcome: Model 1 included self-reported verification; Model 2 added mean task confidence; Model 3 repeated Model 2 with the stricter alternative scoring of Question 17. HC3 robust covariance was used because a Breusch–Pagan test indicated heteroskedasticity in Model 2 (p = .039).

**Table 9**

*Summary of Exploratory OLS Regression Models*

| Model | n | $R^2$ | Adj. $R^2$ | HC3 Wald F (df) | p | $\Delta R^2$ |
|---|---|---|---|---|---|---|
| M1: Accuracy ~ verification | 38 | .227 | .205 | 8.42 (1, 36) | .006 | — |
| M2: Accuracy ~ verification + confidence | 38 | .455 | .423 | 18.64 (2, 35) | < .001 | .228 |
| M3: Strict Q17 key; verification + confidence | 38 | .297 | .256 | 10.51 (2, 35) | < .001 | — |

*Note.* The outcome was a bounded, discrete count of correct judgments (0–5). HC3 robust joint Wald F tests are shown. Model 3 reclassifies Question 17 as incorrect. Model comparisons are exploratory; no causal interpretation is warranted.

**Table 10**

*HC3 Regression Coefficients Predicting Practical Accuracy*

| Model | Predictor | B | HC3 SE | β | z | p | 95% CI |
|---|---|---|---|---|---|---|---|
| M1 | Verification practice | .974 | .336 | .476 | 2.90 | .004 | [.316, 1.632] |
| M2 | Verification practice | .619 | .332 | .302 | 1.86 | .063 | [−.032, 1.269] |
| M2 | Mean task confidence | .767 | .276 | .508 | 2.78 | .005 | [.226, 1.308] |
| M3 (strict Q17) | Verification practice | .527 | .206 | .393 | 2.56 | .010 | [.123, .931] |
| M3 (strict Q17) | Mean task confidence | .263 | .180 | .266 | 1.47 | .143 | [−.089, .615] |

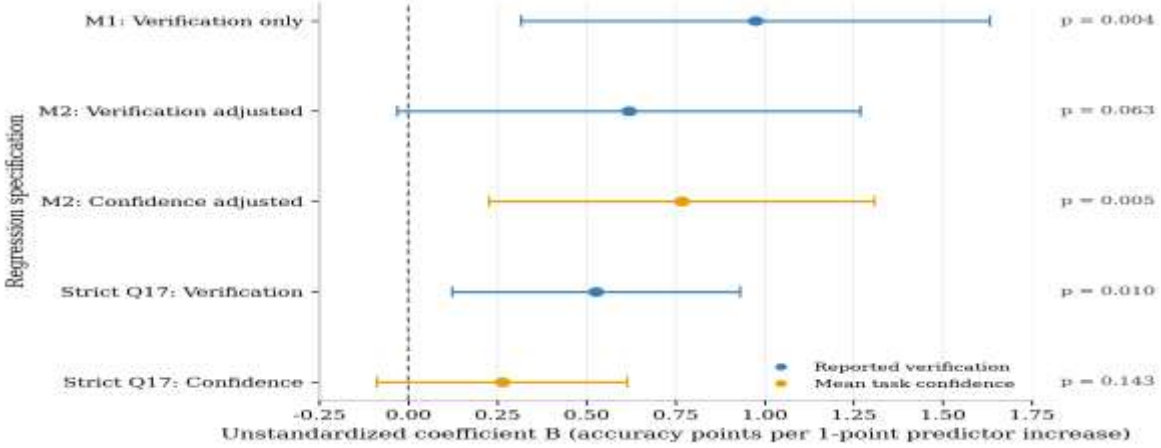
*Note.* B = unstandardized coefficient; β = standardized coefficient; z tests and confidence intervals use HC3 robust covariance. VIF = 1.13 for the two predictors in Models 2 and 3. In Model 2, confidence remained associated with accuracy; the verification coefficient attenuated and did not reach p < .05 after confidence was included. Under the stricter Question 17 key, the verification coefficient remained significant while confidence did not.

The regression results refine the bivariate findings. Verification alone accounted for 22.7% of the sample variance in accuracy ($R^2$ = .227), but its coefficient fell from B = .974 to B

= .619 when confidence was added. In the two-predictor model, confidence was the only coefficient below .05 (B = .767, β = .508, p = .005); verification was positive but uncertain (p = .063). Low VIF values indicate that this attenuation was not due to problematic collinearity. Under the alternative Question 17 key, verification remained associated with accuracy (p = .010), whereas confidence did not (p = .143), demonstrating that conclusions depend partly on task scoring. Figure 5 displays the unstandardized HC3 estimates and their 95% robust confidence intervals.

**Figure 5**

*HC3 Regression Coefficients Predicting Practical Accuracy*




*Note.* Points are unstandardized coefficients B and horizontal bars are HC3 robust 95% confidence intervals; the vertical dashed line is zero. Model 3 uses the strict alternative key for Question 17.

### 4.9 Optional written comments

Thirty-seven respondents supplied a substantive optional comment; one response was recorded as "No answer." The recurring ideas were that AI can produce inaccurate information or language, that students should verify important answers and avoid relying on AI alone, and that students should understand or learn from content rather than copy it. One participant wrote that AI "looks correct but can be inaccurate and lead to plagiarism." These comments complement the closed-response results but were not treated as a formal qualitative dataset with a validated coding scheme.

### 4.10 Discussion by research question

Research Question 1 asked about the level of trust. The four-item mean (M = 3.36) suggests moderate-to-positive reported trust. This does not imply that students consistently accept AI output; the item distributions also show neutral and negative responses. The finding aligns with the view that trust and reliance are related but not equivalent, and that appropriate reliance depends on output and task conditions (Wang, Li, et al., 2025).

Research Question 2 concerned verification. Students reported verification relatively often (M = 3.74), but the practical tasks elicited verification intentions for about two-thirds of judgments. The difference between reported general behavior and responses to concrete examples is consistent with research emphasizing the need to inspect what students actually do, not only what they believe they do. A university case study likewise identified gaps in verification strategies and recommended explicit verification procedures (Martín-Moncunill & Alonso Martínez, 2025).

Research Question 3 asked about judgment accuracy. The 62.1% primary-key accuracy indicates that students often identified the examples correctly, but accuracy varied substantially across content types. The lowest performance on the universal claim about AI grammar suggests that overgeneralized claims about system reliability may be harder to reject than clear grammar errors. Question 17's wording limits the precision of the overall percentage, as the sensitivity analysis demonstrates.

Research Question 4 concerned confidence. Mean confidence was 3.66/5 and was positively associated with accuracy ($\rho = .648$, $p < .001$). In this sample, confidence tended to

track performance at the group level, yet some high-confidence errors remained. The pattern should not be interpreted as proof of complete calibration. A validated performance-based assessment study similarly found that direct performance measures can contribute information beyond self-ratings (Jin et al., 2025).

Research Question 5 concerned AI-use experience and trust. The association between reported use frequency and self-reported trust was positive but did not reach statistical significance ($\rho = .279$, $p = .090$). The small sample and restricted exposure range—37 of 38 respondents reported some use—limit the ability to distinguish experience groups. The available evidence therefore does not show that more frequent use is associated with greater trust in this cohort.

Research Question 6 concerned verification and calibration performance. Higher reported verification was positively associated with accuracy in the bivariate analysis ($\rho = .473$, $p = .003$; Holm $p = .005$ across the three original hypotheses), but this association did not survive Holm adjustment across all 28 exploratory pairs ($p = .071$) and did not remain statistically significant after task confidence was included in the HC3 regression ($p = .063$). The strict alternative scoring of Question 17 also changed which predictor remained significant. These cross-sectional data therefore support a cautious association, not the claim that self-reported verification independently predicts or causes better accuracy.

### 4.11 Chapter summary

The respondents reported moderate-to-positive trust, relatively frequent verification, and moderate practical accuracy. Use frequency was not significantly associated with trust. Verification was positively associated with accuracy in the original bivariate test and remained significant after correction across the three original hypotheses, but it did not survive the broader 28-pair correction and did not uniquely reach significance after controlling for task confidence. Confidence remained positively associated with accuracy. Results are qualified by the achieved sample size, modest trust-scale consistency, one contradictory use-frequency record, the short task set, and sensitivity to the wording of Question 17.

# CHAPTER FIVE
# CONCLUSIONS, AND RECOMMENDATIONS

## 5.1 Introduction

This chapter summarizes the study and its main findings, draws conclusions within the limits of the data, and presents implications and recommendations for EFL teaching and future research.

## 5.2 Summary of the study

The study examined EFL students' reported trust and verification practices and their judgments of five AI-generated English examples at the Faculty of Languages, Surman. A quantitative descriptive-correlational questionnaire was used. The supplied dataset contained 38 responses, below the proposed target of 50–80. Self-report items were analyzed separately from the short performance task, consistent with the distinction between perceived literacy and demonstrated evaluation performance (Jin et al., 2025).

## 5.3 Main findings

- The sample included 38 respondents; 37 reported previous AI use. The largest academic-level groups were second-year (n = 14) and fourth-year (n = 13) students.
- Self-reported trust averaged 3.36/5 ($\alpha = 0.63$), while reported verification averaged 3.74/5 ($\alpha = 0.73$).
- Primary-key performance averaged 3.11/5; 118 of 190 task judgments (62.1%) matched the answer key. Item performance ranged from 47.4% to 89.5%.
- Use frequency was not significantly associated with trust ($\rho = 0.279$, $p = 0.090$).
- Reported verification was positively associated with accuracy in the original bivariate test ($\rho = .473$, $p = .003$; Holm-adjusted across the three original hypotheses $p = .005$). It did not survive Holm adjustment across all 28 exploratory correlations ($p = .071$), and its unique coefficient was not statistically significant after controlling for task confidence ($p = .063$). Mean task confidence was also positively associated with accuracy ($\rho = .648$, $p < .001$; Holm $p = .0003$ across all pairs) and remained significant in the HC3 regression ($\beta = .51$, $p = .005$). Risk awareness (Q9) was associated with adaptive trust (Q14; $\rho = .552$, $p < .001$; Holm $p = .0089$). The two-predictor model explained 45.5% of sample variance ($R^2 = .455$, adjusted $R^2 = .423$).
- Open-ended answers commonly emphasized possible AI errors, verification, avoiding overdependence, and preserving students' own learning.

## 5.4 Conclusions

Within this sample, students appear to hold neither uniformly high nor uniformly low trust in AI-generated English. Their self-reports suggest that they recognize the need for checking, and their practical performance shows a moderate ability to distinguish correct from incorrect examples. At the same time, the incomplete accuracy on several tasks and the presence of confident errors indicate that trust should be connected to evidence and task demands, not to fluent presentation alone.

The positive bivariate association between verification and accuracy is encouraging but requires qualification. It survives correction across the three original hypotheses, not the wider 28-pair exploratory family, and its unique regression coefficient was not statistically significant

after controlling for task confidence. The study cannot determine whether verification improves evaluation skill, whether stronger language knowledge supports both, or whether another factor explains the relationship. Likewise, the group-level confidence–accuracy association does not mean that every confident judgment was correct.

The findings are preliminary and local. They should not be generalized to all Libyan EFL students or used to claim a validated measure of trust calibration. The five-item task is best regarded as an initial performance snapshot.

### 5.5 Implications for EFL instruction

EFL instruction can treat AI responses as drafts or claims to be evaluated rather than as authoritative answers. Students can be asked to identify the precise linguistic or factual claim, compare it with a course reference or trusted source, and explain whether they accept, edit, or reject the output. This approach is consistent with broader AI-literacy work that treats evaluation as a distinct competency (Wang, Li, et al., 2025).

Teachers may model a brief verification routine: isolate a claim; check a reliable source; compare the AI response with course content and context; then document what was changed and why. Tasks should include plausible but incorrect grammar explanations, inaccurate translations, contextual vocabulary problems, and factual statements. Students should also be asked to record confidence before seeing feedback so that confidence–accuracy gaps can be discussed.

The findings do not support prohibiting AI use on the basis of this sample. Instead, they support guided, transparent use with explicit verification expectations. Any institutional policy should distinguish permissible language support from submitting unverified or misrepresented AI-generated work.

### 5.6 Recommendations

1. Increase the sample in a subsequent administration, especially the underrepresented third-year group, while documenting the sampling procedure and response rate.
2. Revise Question 17 to make the intended distinction unambiguous. Have two or three EFL or language-assessment specialists review every task and retain a documented scoring key before data collection.
3. Pilot the revised questionnaire with students outside the final sample. Record completion time, confusing wording, and item-level feedback.
4. Teach source checking and language verification through short classroom activities using dictionaries, grammar references, course materials, and reliable sources.
5. Separate self-reported trust, self-reported verification, AI-awareness, and practical accuracy in future analyses. Do not combine them into one score without validation.
6. If researchers want to make claims about individual trust calibration, use a larger number of balanced items with confidence judgments and a prespecified scoring approach, and report reliability and validity evidence.
7. Resolve the inconsistency between the "ever used AI" response and reported frequency during data cleaning in future forms, for example by adding form logic that requires Never when the answer is No.

### 5.7 Limitations

The achieved sample of 38 was below the planned 50–80 and was recruited through a voluntary-response process, so selection bias and limited generalizability are possible.

Academic-year group sizes were uneven. Nearly all respondents reported some AI use, leaving little variation for testing use-experience relationships. The four-item trust scale had modest internal consistency, and the four-item verification scale should also be interpreted cautiously despite its higher alpha. Cronbach's alpha is a sample-dependent estimate and does not establish validity.

All self-report measures may be affected by recall or social-desirability bias. The task section contained only five examples and included a paraphrase whose key was not entirely unambiguous. Treating "Not sure" as not correct is a transparent scoring rule but collapses uncertainty with a wrong judgment in the accuracy total; the response category was therefore described separately in the interpretation. No experimental manipulation, longitudinal data, or direct observation of verification behavior was available. The file did not include evidence of expert review, pilot testing, ethics approval, or an institutional sampling frame, so this chapter makes no claim that these procedures were completed.

### 5.8 Suggestions for future research

A larger multi-level study could test whether language proficiency, year of study, AI-use frequency, and verification habits explain differences in performance. A follow-up study could observe students completing verification tasks and compare self-reported intentions with actual source use. Experimental classroom research could evaluate whether explicit verification instruction improves accuracy and confidence calibration over time. Future instruments should use multiple clear examples per language domain and establish content validity before they are used to make claims about individual learners.

### 5.9 Final conclusion

The study provides preliminary local evidence that EFL students at the Faculty of Languages, Surman report using verification strategies and show a moderate ability to judge selected AI-generated English content. Verification and confidence were positively associated with practical accuracy, while frequency of AI use was not significantly related to self-reported trust. The educational priority is therefore not simply to increase or reduce trust, but to help students make evidence-based decisions about when to accept, check, revise, or reject AI output.

## References

Alghamdi, L. H., & Alghizzi, T. M. (2026). Integrating synergistic learning and ability grouping with AI-generated multi-tiered tasks to enhance reading attitudes and outcomes. *Acta Psychologica, 263*, Article 106188. https://doi.org/10.1016/j.actpsy.2025.106188

Gill, R. S., Chow, K. M., Dong, H., Wong, H., Fung, M., Myers, P. S., Wong, L. L. C., Lai, P., Ngai, C. S. B., Koon, A. C., Yuen, B., Wong, J., & Jhaveri, A. (2026). Enhancing academic English instruction through pedagogically integrated GenAI: Effects on students' AI literacy and engagement. *Journal of English for Academic Purposes, 83*, Article 101729. https://doi.org/10.1016/j.jeap.2026.101729

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55(12)*, Article 248. https://doi.org/10.1145/3571730

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence, 9*, Article 100436. https://doi.org/10.1016/j.caeai.2025.100436

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46(1)*, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, R., Zhang, C., & Ye, H. (2026). When does GenAI feedback support learning? Trust calibration and verification in L2 writing assessment. *Assessing Writing, 69*, Article 101087. https://doi.org/10.1016/j.asw.2026.101087

Li, T., & Ding, Y. (2025). Exploring the GenAI literacy of Chinese university students in EFL learning. *International Journal of Computer-Assisted Language Learning and Teaching, 15(1)*, 1–24. https://doi.org/10.4018/IJCALLT.377175

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Martín-Moncunill, D., & Alonso Martínez, D. (2025). Students' trust in AI and their verification strategies: A case study at Camilo José Cela University. *Education Sciences, 15(10)*, Article 1307. https://doi.org/10.3390/educsci15101307

Mhamdi, N. A. (2026). AI-mediated feedback in English learning: Moroccan graduate students' accounts of motivation, autonomy, and trust conditions. *Computers and Education Open, 10*, Article 100356. https://doi.org/10.1016/j.caeo.2026.100356

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Wang, F., Li, N., Cheung, A. C. K., & Wong, G. K. W. (2025). In GenAI we trust: An investigation of university students' reliance on and resistance to generative AI in language learning. *International Journal of Educational Technology in Higher Education, 22*, Article 59. https://doi.org/10.1186/s41239-025-00547-9

Wang, Y., Derakhshan, A., & Ghiasvand, F. (2025). EFL teachers' generative artificial intelligence (GenAI) literacy: A scale development and validation study. *System, 133*, Article 103791. https://doi.org/10.1016/j.system.2025.103791

# Appendix A: Questionnaire

Human or Machine? EFL Students' Trust Calibration in AI-Generated English Content

A Study at the Faculty of Languages, Surman, University of Sabratha

Dear Student,

This questionnaire is part of an academic study examining how EFL students use, evaluate, and place trust in AI-generated English content. Participation is voluntary. Your responses are anonymous and will be used only for academic research. There are no right or wrong answers in the opinion section. Please answer honestly based on your actual experience.

Estimated completion time: 8–12 minutes.

**Section A: Background Information**

1. Academic level: ☐ First year ☐ Second year ☐ Third year ☐ Fourth year

2. Have you used AI tools such as ChatGPT for English learning or academic work? ☐ Yes ☐ No

3. How often do you use AI tools for English learning or academic work? ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Very often

4. What do you mainly use AI for? (Choose all that apply.) ☐ Writing ☐ Grammar ☐ Vocabulary ☐ Translation ☐ Speaking ☐ Summarizing ☐ Other: _____

**Section B: Trust, Verification, and Critical Evaluation**

For Questions 5–14, select one response:

1 = Strongly Disagree 2 = Disagree 3 = Neither Agree nor Disagree 4 = Agree 5 = Strongly Agree

5. I generally trust AI-generated English explanations.

6. I believe that AI-generated English answers are usually accurate.

7. I feel confident using AI-generated English content for learning.

8. When an AI answer is written clearly and confidently, I tend to believe it.

9. I know that AI can produce an incorrect answer even when the English sounds natural.

10. I check AI-generated information before using it in academic work.

11. I compare AI-generated answers with textbooks, dictionaries, or reliable sources.

12. I verify AI-generated translations when accuracy is important.

13. I ask a teacher or another reliable source when I am uncertain about an AI-generated answer.

14. I change my level of trust depending on the type of task and the importance of accuracy.

**Section C: Practical Evaluation of AI-Generated English Content**

For each example, please answer A, B, and C. Do not try to identify whether the example was written by AI. Focus on whether the content itself is correct.

A. Is the content correct? ☐ Yes ☐ No ☐ Not sure

B. How confident are you in your answer? 1 = Not confident at all 2 = Slightly confident 3 = Moderately confident 4 = Very confident 5 = Extremely confident

C. Would you verify this answer before using it? ☐ Yes ☐ No

**15. Grammar Example**

AI-generated answer: "I have visited London last year, and I really enjoyed the trip."

Task: Evaluate the English sentence.

A. Correct? ☐ Yes ☐ No ☐ Not sure

B. Confidence (1–5): ☐1 ☐2 ☐3 ☐4 ☐5

C. Would you verify it? ☐ Yes ☐ No

**16. Vocabulary/Usage Example**

AI-generated answer: "I am very interesting in learning English because it helps me communicate with people from different countries."

Task: Evaluate the English sentence.

A. Correct? ☐ Yes ☐ No ☐ Not sure

B. Confidence (1–5): ☐1 ☐2 ☐3 ☐4 ☐5

C. Would you verify it? ☐ Yes ☐ No

**17. Translation Example**

AI-generated answer: "She is looking forward to meeting you" means "She is waiting with excitement to meet you in the future."

Task: Evaluate the explanation of the expression.

A. Correct? ☐ Yes ☐ No ☐ Not sure

B. Confidence (1–5): ☐1 ☐2 ☐3 ☐4 ☐5

C. Would you verify it? ☐ Yes ☐ No

**18. Factual/Language Example**

AI-generated answer: "The past tense of 'go' is 'goed' because regular verbs form the past by adding -ed."

Task: Evaluate the explanation.

A. Correct? ☐ Yes ☐ No ☐ Not sure

B. Confidence (1–5): ☐1 ☐2 ☐3 ☐4 ☐5

C. Would you verify it? ☐ Yes ☐ No

**19. Academic-English Example**

AI-generated answer: "Because AI-generated text is always grammatically perfect, students do not need to check grammar when using it in academic writing."

Task: Evaluate the claim.

A. Correct? ☐ Yes ☐ No ☐ Not sure

B. Confidence (1–5): ☐1 ☐2 ☐3 ☐4 ☐5

C. Would you verify it? ☐ Yes ☐ No

**Question 20: Optional Comment**

20. In your opinion, what is the main reason students should be careful when using AI-generated English content?