Evaluate the influence of LLM on learning of IT students

Goal Question Metrics(GQM)

Software Metrics (SE-611)

Submitted by:

Md. Kamrul hasan: 1418

Md. Nahidur Rahman Nahid: 1429

Nazmus Shakib: 1452

Submitted to:

Dr. Emon Kumar Dey

Associate Professor

Institute of Information Technology

University of Dhaka

IIT

University of Dhaka

Table of Contents

1.Introduction

The GQM (Goal, Question, Metric) approach is a structured technique used to enhance and assess software quality by aligning measurements with well-defined goals. It consists of three key stages:

1. **Goal (Conceptual Level):**
   This stage involves identifying a specific objective that reflects what we aim to achieve. The goal is framed based on the project's context, relevant quality standards, and stakeholder viewpoints.

2. **Question (Operational Level):**
   Once the goal is set, we generate a series of questions aimed at examining various aspects related to the goal. These questions help clarify our focus and ensure we're moving in the right direction.

3. **Metric (Quantitative Level):**
   For every question, we define relevant metrics—measurable values that allow us to collect data and observe trends. These metrics provide the basis for evaluating whether the objective has been reached.

To implement the GQM method, we start by defining a clear goal. Then we develop questions that guide data collection and analysis. Each question is supported by metrics to ensure accurate tracking. In the end, we interpret the collected data to assess our progress toward the goal.

2. Project Specification

2.1 Project Overview

Our objective is to examine how Large Language Models (LLMs) such as ChatGPT, Deepseek, Black box.ai, Gemini,Claude and others are impacting the learning process of IT students. This project aims to evaluate whether the use of LLMs contributes positively to students' understanding, problem-solving ability and academic performance from the learners' perspective.

2.2 Motivation

In recent years, the use of AI tools powered by LLMs has rapidly increased among IT students. These models are capable of explaining complex concepts, solving coding problems, generating documentation, and even assisting in project development. Many students rely on LLMs for learning programming, system design, databases and other technical subjects. While some view this as a powerful aid to learning, others question whether it leads to dependency and shallow understanding about academic honesty, as students may use LLMs to complete assignments without fully engaging in the learning process. These mixed opinions encouraged us to evaluate the real influence of LLMs on IT students' education.

2.3 Scope

This project focuses on students from IT who have used LLMs as a part of their learning activities. We aim to assess how these tools have affected their knowledge acquisition, problem-solving approach, and overall academic experience. Our scope is limited to those students who actively engage with LLMs in their study routine, either for academic tasks or skill development.

3. Goal Specification

3.1 GQM Framework

General Statement: Evaluating the influence of Large Language Models (LLMs) on the learning experience of IT students.

3.2 PPE Approach

Purpose: To evaluate how LLMs are impacting IT students' learning process in terms of understanding, problem-solving and academic performance.
Perspective: To assess the influence from the viewpoint of IT students who use LLMs as part of their academic and self-learning activities.
Environment: The evaluation is conducted within the academic context, where students use LLMs like ChatGPT, Bard, and Claude to support their studies in subjects such as programming, software engineering, and system design.

After specifying our goal through the purpose–perspective–environment approach, our final goal is:
"To evaluate how Large Language Models (LLMs) are influencing the learning experience of IIT students from the students' perspective within the academic environment where such tools are actively used."

3.3 Sub Goals

Our main goal can be divided into the following three sub-goals:

A. Subgoal 1: Understand the students' expectations behind using LLMs

We aim to achieve this subgoal by asking questions like:

- List the main reasons behind using LLM (e.g., ease, speed, clarity).
- Identify the tasks students complete using LLMs (e.g., coding, understanding concepts, documentation).
- Students' consideration on LLMs and peer support.

B. Subgoal 2: Identify the challenges and limitations students face while using LLMs

This subgoal can be addressed through questions such as:

- What percentage of students report situations where LLMs failed to provide adequate help (e.g., during complex project tasks)?
- What challenges do students face when providing effective prompts to get solutions?

C. Subgoal 3: Examine the relationship between LLM usage and students' learning outcome

This subgoal can be addressed through questions such as:

- Do students show overdependence on LLMs?
- Has students' understanding of concepts improved through the use of LLMs?
- Do students feel more confident in solving problems and completing assignments independently?
- Has students' coding ability or debugging skill improved after using LLMs?
- What is the influence of LLMs on students' critical thinking and creativity?

4. Questionnaire Preparation and Data Collection

To effectively evaluate the impact of Large Language Models (LLMs) on the learning process of IIT students, we designed a detailed survey questionnaire aligned with our GQM goals and subgoals. The objective was to gather first-hand insights from students who actively use LLMs in their academic journey.

Questionnaire Design Strategy

Our questionnaire was structured based on the three subgoals derived from the GQM approach:

A. Exploring Expectations

To understand the reasons behind using LLMs, we asked questions such as:

- What type of LLMs (e.g., ChatGPT, Claude, Gemini, DeepSeek, etc.) do students use?
- Which activities are most supported by LLMs? (Options included: Code generation, Concept explanation, Documentation, Debugging, Summarization, Idea generation, etc.)
- What medium do students prefer for quick responses (e.g., terminal commands)?
- What medium do they prefer for detailed explanations (e.g., setup guides)?
- Do students believe LLMs are more helpful than traditional resources (e.g., textbooks, peers, tutors)?

These questions helped us assess the motivations behind LLM usage and the expectations students have from these tools.

B. Identifying Challenges and Limitations

To assess obstacles and limitations, we included questions such as:

- Does the LLM often give the desired answer on the first attempt?
- Have students ever encountered situations where LLMs failed to provide adequate help?

- Do they find it difficult to formulate effective questions/prompts for LLMs?
- Do they face any ethical dilemmas while using LLMs in academic tasks like assignments or lab projects?

These responses revealed usability limitations, ethical considerations, and prompt engineering difficulties that students face while using LLMs.

C. Evaluating Impact on Learning Outcomes

To evaluate the effects of LLM usage on academic performance and learning outcomes, we asked:

- Has your coding ability improved after using LLMs?
- Do you think you understand concepts better with LLM support?
- Are you confident in doing projects (e.g., SPL1, SPL2) without LLMs?
- Are you confident in doing written tasks (e.g., SRS) without LLMs?
- Do you believe LLM usage has reduced your critical thinking ability?
- Would you recommend LLMs to future learners?

These questions helped us quantify the positive and negative educational impacts of LLM usage.

Data Collection Process

- **Tool Used**: Google Forms
- **Timeframe**: The survey was conducted between June 2 and June 5, 2025.
- **Participants**: around 40 students from various IIT batches (including Batch 14, 15, and 16).

- **Data Proof**: Screenshots of form responses, timestamps, and selected anonymous responses have been included in the appendix as evidence.

- **Format**: Most questions used Likert scale ratings, multiple-choice selections, and short answer inputs to ensure both quantitative and qualitative insights.

5. Data Visualization

In this section, the summary of the collected data is represented:

https://docs.google.com/forms/d/19ozldfL9WNjuUSQVzGB-J8rjWOrTaPMaCEbLocpDbzc/edit#responses

6. Metrics Analysis

To evaluate the usage patterns and educational impact of Large Language Models (LLMs) among IIT students, we conducted a comprehensive survey and applied both descriptive and inferential statistical methods. The analysis was structured around themes as task distribution and perceived impact on learning outcomes.

Question 5: Chi-Square Test

Which one do you prefer for detailed responses (e.g., how to set up and run Code::Blocks in Windows)?

- LLM prompting (Asking ChatGPT)
- Web browsing (Search in Brave, Firefox, Google, Edge)

State the Hypotheses

- **Null Hypothesis (H_0):** There is no significant difference in preference between LLM prompting and web browsing.
- **Alternative Hypothesis (H_1):** There is a significant difference in preference between the two methods.

Organize the Data

Method	**Observed (O)**	**Expected (E)**
LLM Prompting	31	19
Web Browsing	7	19
TOTAL	38	38

Expected frequencies (E) are calculated assuming equal preference:
E = Total responses / 2 = 38 / 2 = 19

Calculate Chi-Square (χ^2)

$\chi^2 = \sum(O - E)^2 / E$

$= (31 - 19)^2 / 19 + (7 - 19)^2 / 19$

$= 144 / 19 + 144 / 19$

$= 7.58 + 7.58$

$= 15.16$

Determine Degrees of Freedom (df)

df = Number of categories − 1

= 2 − 1

= 1

Compare with Critical Value

- **Critical χ^2 value at $\alpha = 0.05$ and df = 1:** 3.84
- **Calculated χ^2:** 15.16

Since **15.16 > 3.84**, we reject the null hypothesis.

Reject the null hypothesis.

There is a **statistically significant difference** in preference between LLM prompting and web browsing for detailed responses.

$\chi^2(1) = 15.16, p < 0.05$

Question 8: Z Test

Does LLM give the desired answer on the first try?

- Yes = 15 users (answered correctly on first try)
- No = 23 users

State the Hypotheses

- **Null Hypothesis (H_0):** p = 0.5 (50% users succeed on first try)
- **Alternative Hypothesis (H_1):** p < 0.5 (less than 50% succeed)

This is a **one-tailed** test (testing for *less than* 50%).

Component	Value
Sample proportion ($\hat{p}$)	0.3947
Null value (p_0)	0.5
Standard Error (SE)	0.0811
Numerator ($\hat{p} - p_0$)	-0.1053
Z-statistic	-1.298 (approx)

Reject H_0 if **Z < -1.28**
Our calculated **Z = -1.298**

$-1.298 < -1.28$ –> Reject H0 hypothesis

At 10% significance level, we reject the null hypothesis.

There is **significant evidence at the 10% level** to conclude that *fewer than half of the users get the desired answer on the first try*.

Question 14: **Z Test**

Has your coding ability improved after using LLM?

- Yes = 23 users
- No = 15 users

State the Hypotheses

- **Null Hypothesis (H_0):** p=0.5 (*The proportion of users who say their coding ability improved after using LLM is equal to 50%).*

- **Alternative Hypothesis (H_1):** p>0.5 (*The proportion of users who say their coding ability improved after using LLM is greater than 50%).*

This is a **right-tailed** test.

p^=3823≈0.6053

Component	Value
Null proportion (p_0)	0.5
Sample size (n)	38
Standard Error (SE) SE = $\sqrt{[p_0(1-p_0)/n]}$ = $\sqrt{[(0.5 \times 0.5)/38]}$ = $\sqrt{(0.25/38)}$	0.0811
Sample proportion ($\hat{p}$)	0.6053
Z-test statistic Z = ($\hat{p}$ - p_0)/SE = (0.6053 - 0.5)/0.0811 = 0.1053/0.0811	1.298

Reject H_0 if Z>1.28
Our calculated Z = **1.298**

Since 1.298>1.281 we **reject the null hypothesis**

At 10% significance level, we reject the null hypothesis.

At the 10% significance level, there is statistically significant evidence that more than half of users feel their coding ability improved after using LLM.

Question 15: Z test

Do you think you understand concepts better using LLMs?

Hypotheses

- Null Hypothesis (H_0): No improvement in understanding (true proportion p = 0.5)
- Alternative Hypothesis (H_1): Improvement in understanding (p > 0.5)

Given Data

- Total responses (n): 38
- % reporting improvement ("Yes"): 86.8%
- Number of "Yes" responses (x): 0.868 × 38 ≈ 33

Test Statistic (z-test for proportion)

Component	Value
Sample proportion ($\hat{p}$): $\hat{p} = x / n = 33 / 38$	0.868

Standard Error (SE): $SE = \sqrt{[p_0 \times (1 - p_0) / n]} = \sqrt{[0.5 \times 0.5 / 38]}$	0.081
z-score: $z = (\hat{p} - p_0) / SE = (0.868 - 0.5) / 0.081$	4.54

Critical z-value (α = 0.05, one-tailed):
z(crit) = 1.645

Result
Calculated z (4.54) > Critical z (1.645)

Conclusion:
Reject H_0. There is statistically significant evidence that **LLM use improves conceptual understanding.**

Question 16:

T-test

How much Confidence in doing projects without LLM

Hypotheses

- **Null Hypothesis (H_0):** Students are confident even without LLM ($\mu \geq 3.5$)
- **Alternative Hypothesis (H_1):** Students lack confidence without LLM ($\mu < 3.5$)

Given Data

The survey responses (38 students) are distributed as follows on a scale of 1–5 (1 = least confident, 5 = most confident):

- 1: 5 students (13.2%)
- 2: 12 students (31.6%)
- 3: 15 students (39.5%)
- 4: 5 students (13.2%)
- 5: 1 student (2.6%)

Calculations

Component	**Value**
MEAN (μ) μ = [(1×5)+(2×12)+(3×15)+ (5 + 24 + 45 + 20 + 5) / 38= 99 / 38	**2.605**
STANDARD DEVIATION (σ) Variance (σ²) calculation: [5×(1-2.605)² + 12×(2-2.605)² + 15×(3-2.605)² + 5×(4-2.605)² + 1×(5-2.605)²] / 38	**0.987**
Standard Deviation (σ) √0.987	**0.993**

t-STATISTIC $t = (\mu - \mu_0)/(\sigma/\sqrt{n})$ $= (2.605 - 3)/(0.993/\sqrt{38})$ $= -0.395/0.161$	**-2.453**
CRITICAL VALUE (α=0.05, one-tailed, df=37) From t-table:	**-1.687**

Decision:

Since -2.453 < -1.687, we reject the null hypothesis.

hypothesis at **0.05** significance level.

Conclusion:

There is statistically significant evidence that students lack confidence in doing projects (e.g., SPL1, SPL2, Microservice) without the help of LLMs.

Question 18: Wilcoxon Signed-Rank Test

Has using LLMs reduced your critical thinking skills?

Data Summary (Likert 1-5, n=38)

Response	Frequency	%
1	3	7.9%
2	7	18.4%
3	13	34.2%
4	13	34.2%
5	2	5.3%

State the Hypotheses

- **Null Hypothesis (H_0):** LLM usage may reduce critical thinking skills. (median=3)
- **Alternative Hypothesis (H_1):** LLM usage may not reduce critical thinking skills. (median<3)

Objective: Test if median response > neutral (3) to assess whether LLM usage may reduce critical thinking skills.

Observed Median: 3.5 (between 3 and 4)

Hypothesized Median (H0): 3

Wilcoxon Test Steps

Ranking Absolute Differences from 3:

- Exclude zeros (responses = 3).
- Assign ranks to $|x_i - 3|$ (average ranks for ties).

Summation of Ranks:

- Positive Ranks (W^+): Responses > 3 (4, 5) → Sum = 191.5
- Negative Ranks (W^-): Responses < 3 (1, 2) → Sum = 156

Test Statistic (W): $\min(W^+, W^-) = 156$

Z-Score & p-value:

p-value (one-tailed): 0.39

Result

- $p = 0.39 > \alpha\ (0.05)$ → Fail to reject H_0.
- No significant evidence that median > 3.

Reframed Conclusion for Practical Significance

While statistically insignificant, the data shows:

- 68.4% (26/38) rated ≥3 ("neutral" to "reduced").
- 39.5% (15/38) rated ≥4 ("moderately" or "extremely" reduced).

Interpretation:

- Users perceive a reduction in critical thinking, but the effect is not strong enough to be statistically conclusive with this sample size.

LLM usage may be reducing critical thinking skills" is *partially supported*

- Statistically: No significant evidence (p=0.39).
- Practically: 39.5% report moderate/strong reduction, suggesting a trend worth monitoring.

While not statistically significant (p=0.39), 39.5% of users reported reduced critical thinking skills after LLM usage.

Question 19: Z test

Are you overly dependent on LLMs (e.g., ChatGPT)?*

- Yes
- No

Hypotheses:

- H_0: p = 0.5 (No over-dependence).
- H_1: p > 0.5 (Over-dependence exists).

Given Data (n=38):

- Yes (Dependent): 44.7% (17 students)
- No (Not dependent): 55.3% (21 students)

Test Statistic:

Component	**Value**
Sample Proportion ($\hat{p}$)	**0.447**
Standard Error (SE)	**0.081**
z-STATISTIC $z = (\hat{p} - p_0)/SE$	**-0.65**
CRITICAL VALUE (α=0.05, one-tailed)	**1.645**

Conclusion:

- z (-0.65) < z_crit (1.645). Fail to reject H_0.
- No significant evidence of over-dependence (44.7% "Yes" ≤ 50%).

Question 20: Chi-Square Test

Do you think LLM can replace traditional learning methods(e.g. text books, Peer Discussion, Class lecture etc.)

Null Hypothesis (H_0):

Most learners believe LLMs can replace traditional methods

Alternative Hypothesis (H_1):

Most learners do not believe LLMs can replace traditional methods.

Observed Counts

- Yes = 9
- No = 14
- Maybe = 15
- Total = 38

Yes (60% of 38): 0.60 × 38 = 22.8

No (20% of 38): 0.20 × 38 = 7.6

Maybe (20% of 38): 0.20 × 38 = 7.6

Response	Observed (O)	Expected (E)

Yes	**9**	**22.8**
No	**14**	**7.6**
Maybe	**15**	**7.6**

Chi-Square Test Statistic Calculation

Component	Calculation	Value
Category 1	**$(9 - 22.8)^2 / 22.8$**	**8.35**
Category 2	**$(14 - 7.6)^2 / 7.6$**	**5.39**
Category 3	**$(15 - 7.6)^2 / 7.6$**	**7.21**
Total χ^2	**8.35 + 5.39 + 7.21**	**20.95**
Degrees of Freedom	**3 categories - 1**	**2**

$\chi 2 \approx 8.35+5.39+7.21=20.95$
df=number of categories−1=3−1=2

Critical Value and p-value

- At 5% significance level, critical value (df=2) ≈ 5.99
- Our test statistic: 20.95 > 5.99

  p-value will be very small (< 0.001)

  Reject the Null Hypothesis.

There is strong evidence that the observed responses do not match the hypothesis that most learners believe LLMs can replace traditional methods.

Learners overall do *not* see LLMs as a clear majority replacement.
They appear more split, with many choosing “No” or “Maybe.”

Question 21: T-test

You would recommend using LLMs to future learners"?

Hypotheses

- **Null Hypothesis (H_0):** Mean rating ≤ 3 (neutral or would not recommend).
- **Alternative Hypothesis (H_1)**: Mean rating > 3 (would recommend).

Given Data

- Sample size (n): 38
- Mean rating (μ): 3.63
- Rating distribution:
    - 1: 2.6% (1 student)
    - 2: 10.5% (4 students)
    - 3: 34.2% (13 students)
    - 4: 26.3% (10 students)
    - 5: 26.3% (10 students)

Calculations

Component	**Value**
Standard Deviation (σ)	1.06
T-statistic	3.66

Critical t-value (α=0.05, df=37, one-tailed)	$t(crit) = 1.687$

Conclusion

- Calculated t (3.66) > Critical t (1.687).
- Reject H_0. The data suggests learners highly recommend LLMs (mean 3.63 > 3)

7. Result of Analysis

Result of Analysis

We have a set of 21 questions related to the use of Large Language Models (LLMs) in learning. Among them, we applied Z-tests, t-tests, and chi-square tests to assess hypotheses. The summary of results is shown below:

Question	Null Hypothesis	Test	Result	Conclusion
Q5: Preferred LLM for detailed responses	H0: No significant preference among LLMs for detailed answers	Chi-square	Null rejected	statistically significant difference in preference between LLM prompting and web browsing for detailed responses.
Q8: Does LLM give the desired answer on the first try?	H0: Most users get correct answers on the first try	Z-test	Null rejected	Most learners don't find the first response adequate

Q14: Has coding ability improved using LLM?	H0: improvement in coding skills	Z-test	Null rejected	Users coding skills have improved
Q15: Better conceptual understanding with LLM	H0: No improvement in understanding	Z-test	Null rejected	Conceptual understanding has improved for most users
Q16: Confidence in doing projects without LLM	H0: Students are confident even without LLM	t-test	Null rejected	Students lack confidence without LLM support.
Q18: LLM reduced critical thinking	Hypothesized Median (H_0): 3 (neutral)	Wilcoxon Signed-Rank Test	Null accepted	LLM usage may be reducing critical thinking skills.
Q19: Over-dependence on LLM	H0: Most users are not overly dependent	Z-test	Null rejected	Many learners feel dependent on LLMs
Q20: LLM can replace traditional methods	H0: Most believe LLMs can replace traditional methods	Chi Square-test	Null rejected	Learners see LLMs as supportive but not a full replacement.
Q21: Would recommend LLMs to others	H0: Most would not recommend	t-test	Null rejected	LLMs are highly recommended by learners

8. Conclusion

Summary:
There is a big difference between the preference of using LLM and not. While LLM does provide a challenge to its user, users have improved their conceptual understanding of several topics.
The stereotyping of LLM reducing students' critical thinking skills may not be so correct.
On the other hand, students' coding skills have improved through LLM but students no longer feel confident in their academic tasks without LLM and feel over dependency towards it.
Students don't see LLMs as a replacement of traditional learning either but still recommend new learners for a learning assistant.

Recommendations:
LLMs should be used to gain a conceptual understanding of topics (e.g., the principles of the A* algorithm), rather than for implementing them directly.
LLMs should not be relied upon for routine tasks that users are already capable of completing on their own.
LLMs should be used selectively rather than being applied to every project, assignment, or homework task.
Learners should devote sufficient time to critical thinking and independent problem-solving before seeking assistance from LLMs.
There should be ongoing evaluation of the impact of LLM use on students' learning outcomes.

9. References

SE-611 Course Materials, IIT, University of Dhaka

Software Metrics: A Rigorous and Practical Approach THIRD EDITION by Norman Fenton and James Bieman

Complete questionnaire
https://docs.google.com/forms/d/19ozldfL9WNjuUSQVzGB-J8rjWOrTaPMaCEbLocpDbzc/edit

Data collection responses:
https://docs.google.com/spreadsheets/d/1Hs0jDbnS3i7_9BvAZC7Ac0bWAxl6OeEcsnYLlc3LqO4/edit?resourcekey=&gid=517330200#gid=517330200