Machine vs. Mind: Assessing AI's Ability to Mimic Human Authorship in Fiction

- * Sharath Ananthamurthy, Vice Chancellor, Kuvempu University, Shimoga Dist, Karnataka, India 577451, asharath@gmail.com
- * Sathwik K J, Research Scholar, Department of PG Studies and Research in English, Kuvempu University, Shimoga Dist, Karnataka, India 577451, sathwik@vulcans.co.in
- Rama Prasad B.V., Chairperson, Department of PG Studies and Research in English, Kuvempu University, Shimoga Dist, Karnataka, India 577451
- Rachel Bari, Research Supervisor, Department of PG Studies and Research in English, Kuvempu University, Shimoga Dist, Karnataka, India 577451

Introduction:

Understanding AI's capabilities, limitations, and ethical implications requires focused research, particularly in how humans perceive and interpret AI-generated literary content. AI is being incorporated more and more into a variety of sectors as it develops, such as healthcare, banking, education, and the arts. After the popularisation of Google's celebrated research paper 'Attention is All You Need' (2017), generative AI has taken giant leaps. Research in AI is essential to avoid biases, increase transparency, and enhance performance while promoting responsible growth. Additionally, research encourages creativity, which makes it possible for AI to improve human productivity, solve challenging issues, and advance science. AI research is crucial in creative fields because it helps define the lines between human and machine-generated work and brings up important issues related to authorship, authenticity, and intellectual property.

In his 1950 paper "Computing Machinery and Intelligence," Alan Turing proposed the Turing Test, a standard for assessing a machine's capacity for human-like intelligence (1950). In the test, a human assessor converses in normal language with a computer and a person, both of whom are unaware of which is which. The machine is considered to have passed the test if the evaluator is unable to consistently differentiate it from the human-based solely on responses. Turing's concept marked a revolutionary change in artificial intelligence (AI), shifting the focus from defining "thinking" to behaviour observation.

There have been thousands of instances of Turing tests. Early attempts, like ELIZA by Joseph Weizenbaum in the 1960s, showed that even basic programs might occasionally deceive people by imitating speech patterns (1966). The difficulties of replicating human psychological states were later brought to light by programs like PARRY, which imitated a person with paranoid schizophrenia. Founded in 1990, the Loebner Prize rose to prominence as a whetstone for Turing Test contests, demonstrating the development of chatbot technology. Recently, advancements in large language models (LLMs) have reignited debates about the relevance of the Turing Test. These days, models that operate ChatGPT, Gemini, and other cutting-edge AI systems exhibit such fluency and coherence that it is difficult to tell if they are actually "thinking" or merely producing complex patterns.

Literature Review:

Case studies using literary texts provide unique insights into how AI understands and generates human-like narratives. Researchers from Carnegie Mellon University and the University of Pennsylvania collaborated on a study (2019) where participants read excerpts from both AI-generated and human-written works. The results showed that while AI could mimic stylistic characteristics like thematic coherence and sentence structure, it often struggled with more intricate story features like character development and emotional complexity. A different case study, administered by the University of Pittsburgh using 2400 research subjects, suggested that modern AI can produce texts nearly indistinguishable from famous authors like Emily Dickinson and Shakespeare (2024). Additionally, a Microsoft project (2019) focused exclusively on Shakespearean-style text generation. Even though AI was able to accurately mimic Shakespeare's poetic style, human readers were still able to pick up on minute variations in the emotional resonance and thematic depth. These case studies raise important issues regarding authorship, creativity, and machine intelligence while illuminating the advancements and constraints of AI in literary production.

The growing complexity of AI in producing literature that resembles that of humans and the difficulties in differentiating between machine-generated and human-authored content make Turing Tests based on literary texts necessary. Literary Turing Tests assess AI's capacity to write narratives, poetry, and prose with nuance, coherence, and emotional resonance, in contrast to standard Turing Tests that concentrate on conversational AI. Evaluating AI's ability to create genuine literary works is essential for defining the limits between artificial and human creativity as it gets more and more incorporated into creative disciplines. Furthermore, literary Turing Tests offer an ethical framework for assessing and controlling AI's role in literature and media in light of the growth of AI-generated disinformation and synthetic media. It may raise some serious questions, which might lead to interesting insights.

Despite a large body of prior research on the topic, researchers hypothesised that using a modified Turing Test could provide insightful information. In addition to test replies, researchers also gathered respondents' personal information. The researchers' main goal was to find out if a machine could accurately replicate human writing, but they also wanted to find out if different backgrounds of students affected the outcomes.

Plan and Procedure of the experiment:

The design of the experiment was simple. 52 volunteer subjects from a variety of social backgrounds read two texts: one created by a machine and the other by a professional writer. Subjects were either students of M.A. English or Ph.D. scholars of Kuvempu University, India. Students of the two programs were selected for the study with the assumption that they would be proficient with language and literature enough to evaluate the quality of literary texts. They were not informed ahead of time which text was produced by AI in order to guarantee an objective assessment. Following the reading, they completed a questionnaire consisting of two sections: personal details and eight evaluative questions. Then, researchers used statistical techniques to examine the responses in order to gain insights into their key research questions.

Researchers chose a text they believed would be obscure to the respondents. They reasoned that identification would be too easy if the text were familiar, which could jeopardise the accuracy of the findings. The short story ‘The Third Bank of the River’, originally authored in Portuguese by João Guimarães Rosa (1967), translated to English by

William L Grossman, was chosen for the task. The text was chosen with an assumption that it might be an obscured text for the subjects, so that it doesn't impact the results. Henceforth, this text will be called 'Text–A'. A machine-generated text was produced using the GPT-4 engine by uploading the text of The Third Bank of the River along with the prompt: 'Study the story and rewrite it. (not rephrase) Maintain the depth of the story and emotions, but can change some details.' Generated text will be henceforth called 'Text–B'

The first part of the questionnaire focused on collecting personal information about the subjects. The details gathered included name (optional), gender, program of study, schooling location for Grades 1–7 and Grades 8–10, and the medium of instruction for Grades 1–7 and Grades 8–10. The researchers aimed to determine whether gender influences the perception of literary texts. The participants were registered for either the Ph.D. in English program or the M.A. degree in English. In order to investigate any distinctions between rural and urban schooling, the location of the school was deemed relevant. The medium of instruction referred to the primary language of learning at school, either the native language or English. Certain students studied all major subjects like Maths, Science, and Humanities in English at school. Some students studied them in Kannada (local language). Data for Grades 1-7 and Grades 8-10 were collected separately, recognising the trend of transitioning into new schools after 7th grade.

The second part of the questionnaire contained 8 questions. The first 6 questions focused on evaluating the quality of the two texts. It read 'Here are 6 questions to evaluate the quality of short stories. Respond with a scale of 1 to 5 (1 = Strongly Disagree, 5 = Strongly Agree)'. Six statements applicable to both Text–A and Text–B were, 'Text is clear and well Structured', 'The ideas are logically organized and flow smoothly', ' The vocabulary is appropriate and enhances the readability of the text', 'The text captures and holds the reader's interest.', 'The characters are well-developed and relatable', 'The story evokes strong emotions or thought-provoking ideas'. Question 7 is the main Turing test question: Categorise the two stories into these classifications. (Written and translated by humans, Generated by a computer)'. Question 8 demanded a descriptive response: 'Why did you come to the above conclusion?'

Result:

As seen in Table–1, Respondents gauged the quality of text–A (Written and translated by a human) as follows: For Clarity and structure 3.45, Organisation and flow 3.03, Vocabulary and readability 3.55, Capturing interest 3.37, Characters 3.46, Emotions and Ideas 3.48 on average. For text–B (Generated text), Clarity and structure 3.89, Organisation and flow 3.98, Vocabulary and readability 3.94, Capturing interest 4.13, Characters 3.96, Emotions and Ideas 3.82 were scored on average.

As per the response of the subjects, the quality of the AI text surpasses the written text in all parameters. However, the difference provides us with an interesting insight. When it comes to organisation and flow, and capturing interest, the gap is very significant, denoting that AI is very strong in these areas. Despite AI surpassing written text in Emotions and ideas, the gap isn't significant compared to the other parameters.

For question 6, out of 52 respondents, 18 respondents considered text–A (human) to be written and translated by a human, and 34 considered it to be written by a computer. The

accuracy of the respondents while determining the nature of the text is 34.61%. As there are only two options, usually, if the accuracy is near 50%, the engine is considered to be able to fool people. But what does 34.61% mean? in other words, what does 'computer is able to write like human beings better than humans do' mean? The only sensible explanation would be, that it isn't. However, this indicates that there are some prejudices or biases working behind respondents' minds.

Gender-wise result:

Table–2 demonstrates gender-wise results. Out of 26 male respondents, 5 respondents considered text–A to be written and translated by a human, and 21 considered it to be written by a computer. Thus, only 19.23% of the male respondents were accurate. Out of 26 female respondents, 13 respondents considered text–A to be written and translated by a human, and 13 considered it to be written by a computer indicating the accuracy of female respondents in classifying the text as 50%.

Program of study:

Table 3 demonstrates results based on the program respondents' studying. Out of 13 research scholar respondents, 3 respondents considered text–A to be written and translated by a human, and 10 considered it to be written by a computer. The accuracy of research scholar students in classifying the text was 23.07%. Out of 39 postgraduate students, 15 respondents considered text–A to be written and translated by a human, and 24 considered it to be written by a computer. The accuracy of Post-Graduate Students in classifying the text was 38.46%

Other results:

Data on the medium of learning during primary education was collected. However, this data do not allow us to come to any conclusions with regards to the relation of the medium of instruction and accuracy of the respondents in identifying the texts.

Preliminary observations:

The following are some preliminary observations made:

- As a group accuracy of determining if a test is written by a computer or a human is 34.61%, which is quite below the expected accuracy (close to 50%). It can be inferred that there are biases or prejudices influencing how the texts are being perceived.
- The accuracy of female respondents is 50%, which is the expected level. The accuracy of male respondents is 19.23%, which is way below the expected level.
- The accuracy of research scholars in classifying the text is 23.07%, and the accuracy of post-graduate students in classifying the text is 38.46%. i.e., higher the education level, lower the accuracy. One explanation that seems counterintuitive could be that higher education contributes to the development of biases. However, unless the nature of such biases is known, it can't be conclusive.

Respondents' Reasoning:

Question number 8 of the questionnaire probes for a detailed explanation of the rationale behind their stance. There are some interesting observations to be made. Firstly, a common assumption carried by a large number of respondents is that if the quality of the text

is deemed to be superior, then such a text must have been written by a skilled writer. However, if we refer to Table–1, we can notice that this isn't the case. That is, we surmise the major bias which is contributing to the mistake in identifying texts.

The second observation that can be made is that most responses which identified the text's authorship correctly cited linguistic reasons. By this, we mean the identification of certain words or phrases that the respondents think might not have been produced by a computer, sentence structure or paragraphing. Two of the respondents noticed the usage 'had had', and considered this repetition to be a grammatical error. Though the usage isn't wrong, the conclusion arrived at is right. Contrary to that, a lot of responses which mistook the texts cited literary quality reasons. This might answer why we can notice the difference found when we notice gender-wise data and program of study data. Research scholars might have tried to analyse the text while focusing on literary quality, whereas postgraduate students might have leaned towards language analysis. Further, we looked for the possibility if there is a gender difference in perceiving the language. Two opposite studies can be referred to here. The study led by Iris E. C. Sommer on language lateralisation and sex difference (2003). It failed to conclusively assert any gender difference in language perception. Another study led by Douglas D. Burman identified that the left fusiform region of the brain is, on average, more active among girls than boys, which might lead to better comprehension of language (2008). This study might give us an idea of why the huge difference in gender related data is observed. However, that might not be on the only factor at play here. Socio-economic conditions of rural India might have influenced the results. Poverty is widespread in the region and pressure to earn falls on men in the society that the respondents represent. Male children of the region are often encouraged to pursue degrees that fetch good income. Boys of higher intelligence often get into engineering and medicine compared to girls of similar aptitude. While female children may venture into degree like literature for the sake of recognition or better matrimonial prospects. I.e. it is possible that the female respondents of the experiments might be more academically inclined than the male respondents. So we suspect that the balance is tipped towards one side influencing the result.

Comments on the credibility of the experiment:

We provide a few comments on the limitations of our study, wherein some caution may be in place while drawing conclusions. Firstly, the sample size was limited and small as the experiment was conducted with students from a single discipline, namely, English. Thus, while speculating regarding the biases, we observed that we are unable at this stage to attribute "universality" to our findings. Only two texts were used for the experiment. Our findings would have been more robust had we experimented with more than two texts, and with a distribution of a variety of genres representing pairs of texts, each pair consisting of a machine-generated and a human-generated text. The subjects of the experiments were non native speakers of English. This might have influenced the perception of the text. For these reasons, quantitative data gained by the experiment may be only an indication of possible trends in the perception of texts written and generated. However, this exercise can lead to some insights which can be explored in further research.

Discussion:

The first observation made is that the writing generated by the machine is perceived to be of superior quality to that written by a creative writer. This is consistent with the recent

research findings mentioned in the literature review. As the datasets that fuel the AI engines increase, and optimize the quality of the output keeps getting better. Secondly, AI writing is presumed to be human writing by a larger group and vice versa. This is also consistent with many recent attempts at Turing tests. Even though the respondents of this experiment are non-native speakers of English, other experiments conducted in different parts of the world have produced similar results. The experiment conducted by us involved only 52 subjects; hence, the result isn't conclusive. However, it is reassuring to note that a University of Pittsburgh study, which involved 2400 subjects, also produced similar results. The Turingbench study and Microsoft research reflect the same. It is indisputable that AI engines have passed the milestone of imitating humans. The Loebner Prize, a popular yearly prize awarded to the most human-like chatbot, has been discontinued since 2020, as the technology has surpassed the threshold of mimicking humans in writing.

This raises an important question. 'Is the Turing test not relevant anymore?' The Turing Test, which was formerly used as a standard to assess artificial intelligence, is no longer a useful indicator of machine intelligence. Since AI has advanced to the point where models can now produce human-like responses without true understanding or reasoning, the test that was first created to determine whether a machine could realistically mimic human discourse is no longer relevant. Although large language models are capable of passing the Turing Test in many situations, they are not capable of independent cognition, self-awareness, or actual comprehension. Furthermore, AI is currently very good at specific jobs like data processing, medical diagnoses, and creative problem-solving, where passing a conversational test is not very important.

For a very long time, chess was considered to be a gauge of intelligence. Computer scientists toiled to achieve a computer which can defeat humans. It was thought that if a computer could defeat a skilled chess player, then the computer could be considered intelligent. In May 1997, IBM's Deep Blue defeated the then-world champion Garry Kasparov. This was a great moment for computers. However, scientists didn't consider this a moment when the computer is intelligent. But the definition of machine intelligence no longer relies on it playing chess. Similarly, at the moment, computers can write as well as any skilled writer, but that doesn't prove their intelligence. It can be confidently asserted that Turing tests can no longer be used to evaluate the intelligence of computers. Does this mean conducting Turing Tests should come to an end? The result of the experiment conducted during this research gives us an insight into the possibility of its other utility. We may no longer be able to assess the intelligence of computer engines using Turing tests, but we may learn a few things about our perception of literary text. In simple terms, the Turing test may not give insight into computers, but it can give insight into humans when they interact with a machine-created piece of work. We suggest that the Turing test, in the manner seen above, can be used in different ways.

Observations on Education:

There are two major observations made in the study that make us rethink higher education. Firstly, while the subjects are from the English Literature department, they are found to be prone to biases regarding the texts. A couple of responses to the question number indicate that it might be because of a lack of a strong understanding of Grammar. Being from

a rural area, students are less exposed to English usage, though their field of study is English. We have to rethink the strategies we use in English education in areas where the English presence is minimal, such as villages in developing countries. Secondly, the results of the study point to the possibility of higher education strengthening the biases rather than erasing them. This is an alarming issue that needs to be addressed.

Recommendations:

Based on the insights gained through the research, the researchers recommend the following:

- The experiments shall be replicated under different scenarios to strengthen the validation of the results. Experiments should include a large number of subjects. And the subjects should contain more individuals from even more diverse backgrounds.
- Turing tests should focus on learning about humans, along with learning about computers.
- The question of whether women are more inclined towards language proficiency shall be probed even more.
- Influence of gender on the choice of program of study can be further studied
- More tests should be conducted to test whether rural/urban origin and medium of instruction influence the perception of the texts.
- The findings raise an even more important question. If computers can mimic human fiction writing, there must be elements in the very nature of writing that are replicable. The nature of writing can be further studied for better insights.
- The method of teaching English in remote areas needs to be rethought.
- One of the observations made in this research is that there is a possibility that the respondents with a higher level of education may be prone to biases much more than the others. This raises interesting questions and issues about the nature of perceiving texts and the biases that come with gaining a higher level of education.
- Since AI is rapidly expanding and increasingly influential, understanding its potential and limitations can be integrated into formal higher education curricula.

Conclusion:

In conclusion, we have tried to expand the possibilities of the Turing test and use this to gain insights about certain biases that accompany higher education. The findings from this research underscore the evolving capabilities of AI in generating human-like text while also revealing the challenges in distinguishing between human and machine-authored writing. The insights gained through this study reinforce the need for continued research on AI's role in creative writing, its limitations, and its implications for authorship and literary authenticity. The research concludes that while the traditional Turing Test as a measure of AI intelligence might have diminished in significance, its utility remains critical in studying human biases, perceptions, and the deeper cognitive frameworks underlying literary analysis. The research also raises some questions about the nature of education and how it might influence the formation of biases.

References:

A, S., & John. (2015). Survey on Chatbot Design Techniques in Speech Conversation Systems. International Journal of Advanced Computer Science and Applications, 6(7). https://doi.org/10.14569/ijacsa.2015.060712.

Burman, D. D., Bitan, T., & Booth, J. R. (2008). Sex differences in neural processing of language among children. Neuropsychologia, 46(5), 1349–1362. https://doi.org/10.1016/j.neuropsychologia.2007.12.021.

Chess.com Team & Chess.com Team (CHESScom). (2018, October 12). Kasparov vs. Deep Blue | The Match That Changed History. Chess.com. https://www.chess.com/article/view/deep-blue-kasparov-chess#kasparov-deep-blue-1997-rematch.

Kuipers, B., McCarthy, J., & Weizenbaum, J. (1976b). Computer power and human reason. ACM SIGART Bulletin, 58, 4–13. https://doi.org/10.1145/1045264.1045265.

OpenAI. (2024). ChatGPT (Dec 15 version) [Large language model]. https://chat.openai.com

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-76900-1.

Rosa, J. G. (1967). The third bank of the river. In University of California Press eBooks (pp. 125–130). https://doi.org/10.1525/9780520312425-015.

Sommer, I., Aleman, A., & Kahn, R. (2003). Do women really have more bilateral language representation than men? Results from a meta-analysis of functional imaging studies. Schizophrenia Research, 60(1), 233. https://doi.org/10.1016/s0920-9964(03)81225-2.

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236), 433–460. https://doi.org/10.1093/mind/lix.236.433.

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TURINGBENCH: A benchmark environment for Turing Test in the age of neural text generation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (2021st ed.). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.172.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168.

Writer, S. (2019, April 30). Machines can't quite crack Shakespeare. Panama City News Herald. https://www.newsherald.com/story/opinion/editorials/2019/04/30/guest-view-machines-cant-quite-crack-shakespeare/5308960007.

Tables:

Table-1

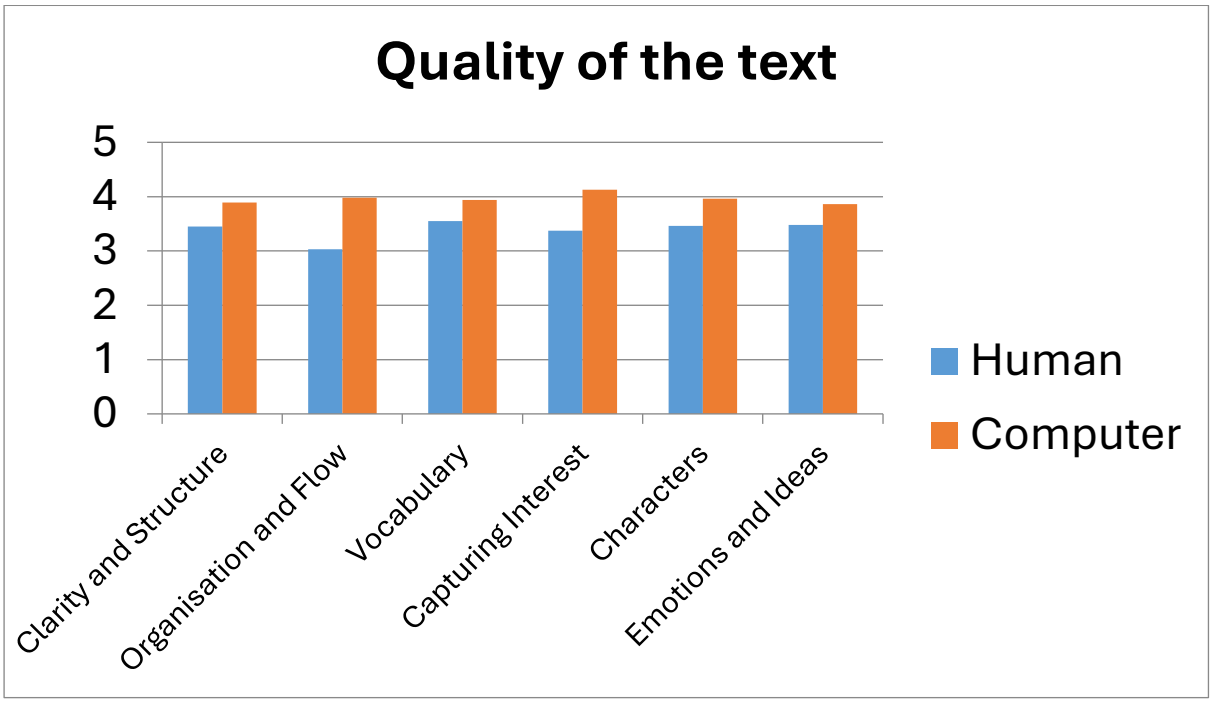


Table-2

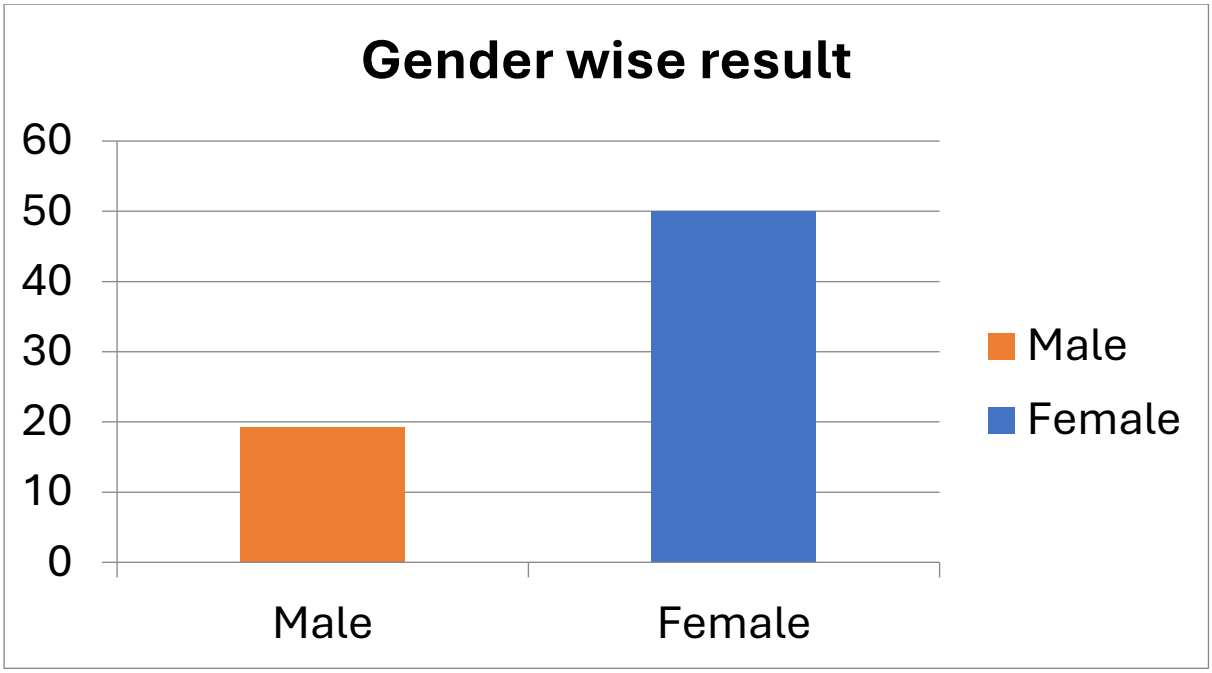
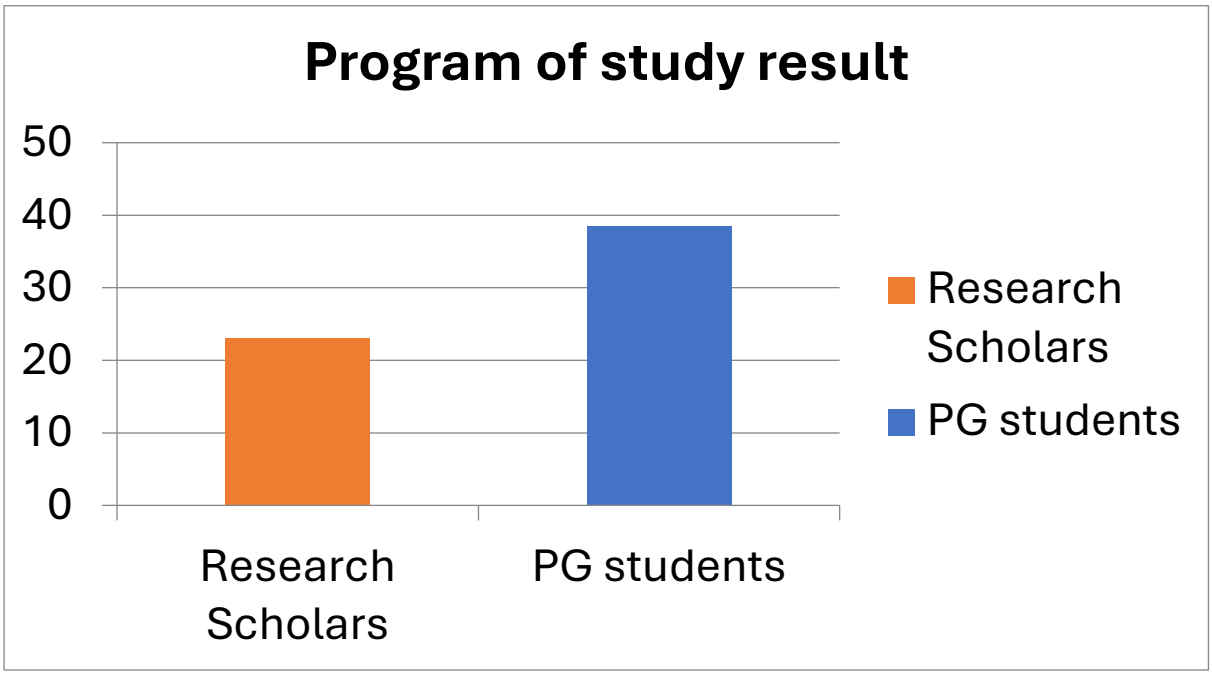

Gender wise result
60
50
40
30
20
10
0
Male
Female
Male
Female

Table-3


Program of study result
50
40
30
20
10
0
Research Scholars
PG students
Research Scholars
PG students