Contents

Contents 1

1.0 Introduction 1

2.0 Literature Review 2

2.1 Rise of the Large Language Model (LLM) 3

3.0 Methods 4

4.0 Results 5

5.0 Summary 6

References 7

1.0 Introduction

This page is intentionally blank.

2.0 Literature Review

The following literature review presents the necessary background to better understand Large Language Models (LLMs) and their impacts on how novice programmers learn to code. We begin by exploring key research around LLMs and how they have impacted disciplines since their inception. Specifically, we provide evidence of LLMs as a disruptive innovation in the field of computer programming. LLMs have transformed not only how people code, but the ways in which programmers think about programming itself. This has implications for not only how code will be written in the future, but what skills are necessary for becoming a programmer and the methods by which those skills are taught to novices.

In addition, this literature review will explore the research of LLM use by teachers, students and administration within higher education. The research embodying the strengths, drawbacks, and unique challenges of LLMs within this context is discussed. This is necessary to differentiate between the overarching issues of LLMs in education and those specific to the topic of computer programming education.

Research will show LLMs are transforming how skilled professionals write code. LLMs are being used predominantly by teachers and students in computer programming education as well, which has forced academics to rethink not only how to teach computer programming to novices, but what should be taught in the first place. The literature will identify this through the positive and negative impacts LLMs are having programming education.

The academic literature on computer science education discusses the well-known challenges of novice users learning to program. These challenges are the impacts of cognitive load on learning, getting learners to focus on computational literacy over syntax and technology, the role of self-efficacy and motivation on success, and the importance of metacognition for building problem-solving skills. This literature review will revisit each of these challenges through the lens of how LLMs are impacting them both positively and negatively.

2.1 Rise of the Large Language Model (LLM)

Natural Language Processing (NLP) has undergone significant changes in the past few years. With the introduction of transformer-based neural network architecture, Natural Language Processing took a giant step forward in performance and accuracy (Gillioz et al., 2020).

The Generative Pre-Trained Transformer (GPT) reasonably predicts the next word in a sequence using the input and previously generated output. While predicting the next token in the sequence is the extent of their ability (Shanahan, 2024), transformer-based language models trained on large data have demonstrated highly effective reasoning capabilities.

Open AI's foundational paper, "Language Models are Few-Shot Learners", demonstrated these transformer-based language models can produce human-level performance when they are trained on large copra (Brown et al., 2020). This was a pivotal discovery because at the time since prior to this paper transformer-based models were trained to be relatively task specific (Zhao et al., 2023).

The paper from Brown et al., 2020 led to significant advancements in research with respect to understanding the capabilities of LLMs. Wei et al. (2022) discovered reasoning capabilities of LLM's can be improved through a technique called chain-of-thought prompting. By including few-shot examples that break down complex reasoning into steps, the LLM can use those shots provided as an example of how to explain a complex process.

Halevy et al's.(2009) seminal paper, "The unreasonable effectiveness of data", explains as the training data set size increases, so does the model accuracy. In addition, the specific selected model algorithm becomes less relevant as training data set size increases. (Wei, Bosma, et al., 2022) documented a similar effect with large language models. Larger-sized models exhibited emergent abilities not found in their smaller counterparts. Examples of emergent abilities seen in the larger models include complex arithmetic and reading comprehension.

2.2 LLMs as a Disruptive Innovation for software development

Christensen et al., (2018) divide technological innovations into two distinct types. Sustaining innovations improve existing products and services, while disruptive innovations provide a unique set of features to an initial set of customers. From the perspective of companies that offer generative AI such as Google,

Anthropic, Open AI, and Microsoft, Horn considers generative AI to be a sustaining innovation (Horn, 2024). In their comprehensive literature review of AI as a disruptive innovation, Păvăloaia & Necula, (2023) consider the application of Generative AI to be a disruptive innovation across different sectors such as healthcare, agriculture business and education.

The rise of AI-assistant programming tools in industry is evidence of this disruption. There are a growing set of tools available in the cloud: Github Copilot, Amazon CodeWhisperer, Gemini Code assist, Claude Code, Open AI Codex v2, Tabnine, and Codeium, in addition to self-hosted options like FauxPilot, Tabby, and CodeLLama. While each provides a unique set of features for differentiation, they all have primary functions like code completion, code generation, code explanations, and discussion. The primary value-add touted by these tools is developers will be able to write code in less time, improving productivity. Talk about "vibe coding" and Lovable, Bolt, Replit and Cursor.

3.0 Methods

3.1 Introduction

The objective of this research is to study the impact of generative AI on an individual's learning. Specifically, this research will study the impacts of Large-Language Model (LLM) use by students enrolled in an introductory Python programming course. Their learning will be studied through the lenses of academic performance on a summative assessment at the mid-term of the course in addition to scores on a computational literacy instrument. LLM use will be classified into activities that either support or avoid learning, which will then be quantified by for each participant. This chapter explains the methodology that will be used to address the research questions of this study.

3.1.1 Research Questions

RQ1: How does large-language model use influence student learning performance?

- This was measured by first classifying chat logs into counts of const/unconst then

- performing a multi-variate regression of the key influencers Constructive/Unconstructive => E1

RQ2: When the prompt is adjusted to include assignment instructions (in-context learning) what is the impact on student learning performance?

- This was measured by first classifying chat logs into counts of const/unconst then

- Checking control / treatment against E1 for any statistical significance.

RQ3: What is the relationship between large language model use and computational literacy?

- This was measured by first classifying chat logs into counts of const/unconst then

- performing a multi-variate regression of the key influencers Constructive/Unconstructive => E1

3.2 Study Design

A diagram of the study design
Figure1: An overview of the study design.

3.2.1 Overview

The study took place over a six-week period in the Spring 2025 semester of an introductory Python programming course, IST256 https://ist256.com/spring2025/syllabus/. Taught within the School of Information Studies at Syracuse University, the course teaches programming fundamentals from the informatics perspective and is intended for non-computer science majors. There were 173 students enrolled in the course. The study only focused on the first 6 weeks because these units cover basic computational literacy constructs as applied to the Python programming language. These include instruction sequencing, variables, branching, iteration, and composition (functions). The course schedule can be found here: https://ist256.com/spring2025/syllabus/#course-schedule. The study was exempt from IRB under section §46.104 section 1, which covers research conducted in an established educational setting. A university IRB Exemption application has been filed and obtained for category 1 for research in established or commonly accepted educational settings. The IRB# is 24-346. While all elements of the study design were part of the course, students could elect to opt-out of including their data in the study.

The study begins with a diagnostic instrument C1 to get the baseline of computational literacy for each participant. This happened within the first week of class before any instruction. At this time students were introduced to the course-provided AI, https://ai.ist256.com, which was a Large Language Model (GPT4o-mini) configured with a system prompt. Students were encouraged to use the AI as a virtual tutor asking it for help with Python questions and course-related assignments. When asking a specific question about an assignment, students were instructed to switch the LLM context by selecting the assignment in question from a drop-down menu.

Figure 2: Context-Selection from the IST256 AI Tutor

For the control group T1 this action did nothing - it does not add any additional context. For the treatment group T2 the action copied the assignment or lab instructions into the conversational context.

Figure 3: The treatment group (T2) is aware of the selected content.

Because the chatbot was self-hosted, all student interactions and AI responses were captured. D1 chatbot trace data is a dataset of those collected interactions throughout the six weeks of use. After the six week period, there were three observations. E1 was the midterm exam in the course covering the content from the first six weeks, C2 was a re-issue of the same C1 diagnostic as a means to measure improvement of computational literacy. Observation Q1 was a questionnaire that asked for simple demographic data about each participant. Q1 was issued at the end of the course along with course evaluations.

3.2.1.1 How the Study Addressed the Research Questions All three research questions explored the influence of Large-Language Model use on learning and computational literacy. Therefore an analysis of the chatbot trace data D1, was crucial to answering the research questions. Categorical content analysis [CITE - foundation content analysis] was used to classify student AI interactions into two categories of learning-supportive and learning-avoidance activities. These activities were deductively coded based on observations of other researchers in literature in addition to my six years of experience teaching the course.[CITE deductive]. The resulting categories were quantified and grouped to establish a quantized profile of AI use by each participant. These category counts were the independent variables of this study.

To answer RQ1, a multivariate regression was performed with the category counts as the independent variables and E1 as the dependent variable. RQ2 introduces the control / treatment independent variable to the multivariate regression from RQ1. Finally for RQ3 the dependent variable is changed to C2.

3.2.2 Participant Funnel

Participants were students enrolled in the Spring 2025 section of IST256. There were 173 students enrolled. The research was exempt from IRB under section §46.104 section 1, which covers research conducted in an established educational setting. A university IRB Exemption application was filed and approved for category 1 for research in established or commonly accepted educational settings under the Syracuse University IRB number #24-346.

Of the 173 participants only 126 consented to the study. Five of the students who consented did not complete the observations necessary for the dependent variables: C1, C2 or E1, leaving 121 participants.

3.2.2.1 Population 1: Chatbot participants There are two populations under analysis. The first population is the set of participants who used the AI chatbot and therefore have the trace data D1 necessary to study their AI interactions. There were 48 individuals in the control

group and 39 in the treatment group for a total participant population of 87. *Figure 4: The participant funnel for chatbot use. 87 participants.*

3.2.2.1 Population 2: Chatbot participants with Survey Responses A survey Q1 was issued to students with the goal of identifying possible covariates. Besides demographic questions around the year of study, major, and gender students were also asked about their programming experience prior to the IST256 course. Ten chatbot users did not complete this survey thus reducing the participant size down to 77 whenever survey responses were needed in the analysis. Among the 77 participants, 41 were in the control group and 36 were in the treatment group.

Figure 4: The participant funnel when accounting for survey responses.

3.2.3 Computational Literacy Instrument (C1/C2)

Explain

3.2.4 Chatbot Design

3.2.4.1 Random Assignment Jxcshgfkjds

3.2.4.2 Base Model Selection and Control Group (X1) dshfgjhsdf

3.2.4.3 Treatment Group (T1) jhgkdfs

3.2.5 Midterm Exam (E1)

Talk about format, questions, closed book, etc..

3.2.6 Questionnaire (Q1)

Jdsfhgkjds

3.2.7 Trace Data

Whats in it how it is collected and labeled?

3.3 Data Analysis

3.3.1 Overview

High level of the process

3.3.2 Tools

3.3.3 Categorizing Trace Data

Explain what you did here, trying out various large language models for categorzing

4.0 Results

TODO

5.0 Summary

This page is intentionally blank.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., & Henighan, T. (2020). *Language Models are Few-Shot Learners*.

Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive Innovation: An Intellectual History and Directions for Future Research. *Journal of Management Studies*, *55*(7), 1043–1078. https://doi.org/10.1111/joms.12349

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). *Overview of the Transformer-based Models for NLP Tasks*. 179–183. https://doi.org/10.15439/2020F20

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Horn, M. B. (2024, June 3). *What does Disruptive Innovation Theory have to say about AI? - Christensen Institute*. https://www.christenseninstitute.org/blog/what-does-disruptive-innovation-say-about-ai/

Păvăloaia, V.-D., & Necula, S.-C. (2023). Artificial Intelligence as a Disruptive Technology—A Systematic Literature Review. *Electronics*, *12*(5), 1102. https://doi.org/10.3390/electronics12051102

Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners* (No. arXiv:2109.01652). arXiv. http://arxiv.org/abs/2109.01652

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). *A Survey of Large Language Models* (No. arXiv:2303.18223). arXiv. http://arxiv.org/abs/2303.18223