

**DESIGNING A PEDAGOGY-DRIVEN AI VALIDATION FRAMEWORK
FOR INTRODUCTORY PYTHON COURSES**

By

Kelly Kirsten Audrey
200001076

MASTER'S DEGREE
in

MASTER OF INFORMATION TECHNOLOGY
DATA SCIENCE BUSINESS INFORMATICS
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

June 2026
Revision after the Thesis Defense on 4th August 2026

STATEMENT BY THE AUTHOR

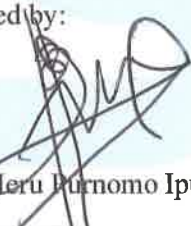
I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.


Kelly Kirsten Audrey

Student

4 August 2026

Approved by:



Dr. Ir. Heru Purnomo Ipung, M.Eng

Thesis Advisor

15 August 2026



Dr. Eka Budiarto, ST, M.Sc

Thesis Co-Advisor

15 August 2026

Dr. Maulahikmah Galinium, S.Kom., M.Sc.

Dean

Kelly Kirsten Audrey

ABSTRACT

DESIGNING A PEDAGOGY-DRIVEN AI VALIDATION FRAMEWORK FOR INTRODUCTORY PYTHON COURSES

By

Kelly Kirsten Audrey

Dr. Ir. Heru Purnomo Ipung, M.Eng, Advisor

Dr. Eka Budiarto, ST, M.Sc, Co-Advisor

SWISS GERMAN UNIVERSITY

The rapid integration of generative AI into programming education has weakened the link between syntactic correctness and genuine understanding, as novices submit Python code using constructs far beyond their curriculum stage, an abstraction leap. This research develops a pedagogy-driven AI validation framework, a proof-of-concept web application named PyGatekeeper, that preserves conceptual learning while permitting responsible AI use. It integrates three pillars: a curriculum ontology mapping 38 Python Abstract Syntax Tree (AST) constructs onto a 16-week course; a deterministic AST gatekeeper detecting out-of-scope constructs; and a Retrieval-Augmented Generation (RAG) engine generating textbook-grounded Socratic questions via Hypothetical Document Embedding and cross-encoder re-ranking. Across three evaluation indicators, the AST gatekeeper reached perfect accuracy (1.00) on 41 submissions; the RAG pipeline met three of four RAGAS targets (faithfulness 0.980); and a live validation with six students and three expert raters showed substantial agreement between the answer-evaluation module and human experts (mean Cohen's $\kappa = 0.730$), confirming the framework's technical feasibility and reliability.

Keywords: Pedagogy-driven AI validation, Abstract Syntax Tree, Curriculum Ontology, Retrieval-Augmented Generation, Socratic questioning



DEDICATION

To my family and beloved one, who never saw education as a burden, but as the greatest gift one can give and receive. This work is yours as much as it is mine.



ACKNOWLEDGEMENTS

First of all I thank God Almighty who has given me the grace and the guidance through each step of my journey. I would like to convey my sincere gratitude to my advisor, Dr. Ir. Heru Purnomo Ipung, M.Eng, for his guidance, knowledge and patience that have made this work what it is today. I would also want to thank my co-advisor, Dr. Eka Budiarto, ST, M.Sc, for his wise insights and support that kept me grounded and focused during the process.

I am also deeply grateful to the participants who made the validation of this work possible. My sincere thanks go to the lecturer and the teaching assistants who generously served as the expert panel, giving their time and their pedagogical judgement to independently review each submission; their expertise is the very foundation of the reliability results reported in this thesis. I am equally thankful to the students who volunteered to test the system, whose submissions and honest feedback turned a prototype into something that could be meaningfully evaluated. This work simply would not have been possible without their willingness to take part.

To my family, dad, mom and brother, thank you for the unconditional love and the faith in education that you placed in me even before I realized what I was reaching for. There's a piece of each of you in this art.

To my beloved one, thank you for the patience, the serenity, the quiet strength you gave me on the days this felt impossible. You made the tough parts easy.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES	10
LIST OF TABLES.....	11
CHAPTER 1 - INTRODUCTION.....	13
1.1. Background	13
1.2. Problem Statement.....	16
1.3. Research Objectives.....	17
1.4. Research Questions.....	18
1.5. Hypothesis	18
1.6. Significance of the Study	18
1.7. Scope and Limitations.....	20
1.8. Thesis Structure.....	22
CHAPTER 2 - LITERATURE REVIEW	24
2.1. Institutional Context	24
2.2. The Introductory Python Course Curriculum	26
2.3. Python Programming.....	29
2.3.1. Python in the Context of this Research.....	29
2.4. Artificial Intelligence in Education	30
2.4.1. Abstract Syntax Tree (AST).....	32
2.5. Curriculum Ontology	32
2.6. Retrieval-Augmented Generation (RAG)	33
2.6.1. Key Retrieval Techniques: Embeddings, Chunking, and HyDE.....	34
2.6.2. Pedagogical and Context-Aware Usage in Education.....	38
2.6.3. Challenges and the Roles of Educators.....	40

2.7.	Detection-Accuracy Metrics (Confusion Matrix)	40
2.8.	RAGAS.....	41
2.9.	Inter-Rater Reliability and Cohen’s Kappa	45
2.10.	Large Language Model (LLM)	48
2.11.	Previous Studies and Critical Comparison.....	49
2.11.1.	Domain 1: Python Fundamental Pedagogy and Course Design	49
2.11.2.	Domain 2: Abstract Syntax Tree (AST) and Code Structural Analysis	53
2.11.3.	Domain 3: Curriculum Ontology and Knowledge Representation	55
2.11.4.	Comparison of Large Language Models (LLMs) for Educational Question and Feedback.....	56
2.11.5.	Domain 4: Retrieval-Augmented Generation (RAG) and RAGAS Evaluation	61
2.11.6.	Domain-Indicator-Source Mapping.....	64
2.11.7.	Expert Panel Qualification Criteria	66
2.11.8.	Synthesis and Selection of Building Blocks.....	68
2.12.	Research Gap.....	72
CHAPTER 3 – RESEARCH METHODS		74
3.1.	Research Framework.....	74
3.1.1.	Research Phases	78
3.1.2.	Mapping of Solution Features to the Identified Problems.....	79
3.1.3.	Pillar 1: Course Pedagogy Design and Curriculum Ontology	81
3.1.4.	Pillar 2: The Pre-Submission Gatekeeper	86
3.1.5.	Pillar 3: Evaluation	91
3.2.	Data Collection Method.....	91
3.3.	Success Criteria.....	91
3.3.1.	Indicator 1: AST Gatekeeper Detection Accuracy	92
3.3.2.	Indicator 2A: RAG Pipeline Quality (RAGAS Framework)	93
3.3.3.	Indicator 2B: Answer Evaluation Reliability (Cohen's Kappa)	94
CHAPTER 4 – RESULTS AND DISCUSSIONS.....		95
4.1.	Proof-of-Concept Overview	95
4.1.1.	Pillar 1: Curriculum Ontology and Pedagogical Framework	96
4.1.2.	Pillar 2: The AST Gatekeeper	98
4.1.3.	Pillar 3: The Pedagogical RAG Engine	98
4.1.4.	End-to-End System Workflow.....	103
4.2.	System Evaluation.....	104
4.2.1.	Indicator 1: AST Gatekeeper Detection Accuracy	104

4.2.2.	Indicator 2A: RAG Pipeline Quality (RAGAS)	107
4.2.3.	Indicator 2B: Answer-Evaluation Reliability (Cohen’s Kappa).....	115
4.3.	Discussion: Addressing the Research Questions and Hypotheses.....	123
CHAPTER 5 – CONCLUSION AND RECOMMENDATION		125
5.1.	Conclusions	125
5.2.	Recommendations	127
GLOSSARY		131
REFERENCES		132
CURRICULUM VITAE		136



LIST OF FIGURES

Figures	Page
Figure 1.1 The usage of AI in Indonesia according to its category	14
Figure 1.2 The usage of AI per category based on generation	14
Figure 1.3 Fishbone Diagram	15
Figure 1.4 Fishbone Diagram with highlighted section	16
Figure 2.1 HyDE query reformulation supported by the curriculum concept map	38
Figure 3.1. (a) Three-Pillar Research Framework, (b) AST Parsing, and (c) Vector and RAG Flow	76
Figure 3.2 Weekly Design and Example	86
Figure 4.1 Field-testing and Focus Group Discussion (FGD) session with student participants, (a) batch 1, (b) batch 2	120
Figure 4.2 Independent validation session with the expert panel (lecturer and teaching assistants), (a) batch 1, (b) batch 2	121

LIST OF TABLES

Table	Page
Table 2.1. Institutional Realities Summary.....	25
Table 2.2. Weekly Topics of the Introductory Python Course	27
Table 2.3. RAGAS Metrics: Definitions, Scope, and Target Thresholds.....	41
Table 2.4. Domain 1 Pedagogical focus	50
Table 2.5. Domain 2 Pedagogical focus	53
Table 2.6. Domain 3 Pedagogical focus	55
Table 2.7. The summarization of the comparison of LLMs	58
Table 2.8. RAG Comparison	62
Table 2.9. Indicators	64
Table 2.10. Profile of Test Person	66
Table 2.11. Cross-Domain Comparison and Selection of Building Blocks	71
Table 3.1. Three-Pillar Research Framework Summary	77
Table 3.2. Research Phases and Their Outputs.....	79
Table 3.3. Mapping of Solution Features to the Identified Problems	79
Table 3.4. Complete CONCEPT_WEEK Curriculum Ontology.....	81
Table 3.5. Pre-Submission Gatekeeper: Four-Stage Technical Implementation	86
Table 3.6. RAG Engine Technical Specifications as Implemented in rag_engine.py ..	88
Table 3.7. Indicator 1: Confusion Matrix Metrics, Formulae, and Targets	92
Table 3.8. Indicator 2A: RAGAS Metrics, Targets, ...	93
Table 4.1. Technology Stack	96
Table 4.2. Indicator 1 - AST Gatekeeper Confusion-Matrix Results	105

Table 4.3. Indicator 2A - RAGAS Results (final HyDE + re-ranking configuration, n = 20, judge = gpt-4o)	107
Table 4.4. Effect of the HyDE + Re-ranking Retrieval Strategy on RAGAS Metrics	108
Table 4.5. Benchmark Comparison of RAGAS Scores Against the Recent Literature	109
Table 4.6. RAGAS Score Progression Across the Four Improvement Phases	110
Table 4.7. Indicator 2B : Cohen's Kappa Between the System Verdict and Each Expert Rater (n = 72 per rater)	117
Table 4.8. Indicator 2B : Inter-Rater Agreement Within the Three-Member Expert Panel (n = 72)	117

