

**AURIVA : ADAPTIVE ENGLISH LANGUAGE
LEARNING PLATFORM FOR AUTISTIC CHILDREN
IN SRI LANKA**

H. H. Jayasundera
(IT22084590)

GNN-Enhanced Adaptive Concept Learning Module
R26-IT-006

Dissertation submitted in partial fulfilment of the requirements for the
B.Sc. (Hons) Degree in Information Technology
Specialising in Information Technology

Department of Information Technology
Sri Lanka Institute of Information Technology
Sri Lanka

October 2026

DECLARATION AND SUPERVISOR CERTIFICATION

Candidate declaration. I declare that this thesis is my own work and does not incorporate, without acknowledgement, material previously submitted for a degree or diploma at any university or institute of higher learning. To the best of my knowledge, it contains no material previously published or written by another person except where acknowledgement is made in the text. I grant the Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute this thesis, in whole or in part, in print, electronic, or other media, while retaining the right to use the content in future work.

Academic integrity and AI disclosure. I have disclosed all use of generative AI and other automated tools in accordance with the module policy, as set out in Appendix B. I remain responsible for the accuracy, originality, citations, code, data, analyses, and conclusions in this thesis.

Student name: H. H. Jayasundera

Student ID: IT22084590

Signature: _____ Date: _____

Supervisor certification. The above candidate carried out this research for the bachelor's degree thesis under my supervision.

Role and name	Signature	Date
Supervisor: Ms. Manori Gamage		
Co-supervisor: Ms. Suriya Kumari		

ABSTRACT

Children with Autism Spectrum Disorder (ASD) in Sri Lanka acquire foundational English vocabulary under conditions that mainstream digital learning tools do not address: an audit of two widely used applications, Otismo and Speech Blub, found that neither provides a tiered mastery architecture, a graduated relearn pathway, teacher-mediated control, or any model of how concepts relate to one another. This thesis presents the Concept Learning Module of Auriva, a teacher-mediated learning platform for autistic children, with its supporting graph intelligence layer. The module teaches curriculum-aligned vocabulary through three tiers of visual familiarization, word recognition, and attribute introduction, governed by a rule-based engine enforcing an Applied Behavior Analysis mastery criterion before advancement. Interactions populate a Neo4j Graph Knowledge Base modelling students, concepts, categories, and observed confusions, over which a Graph Attention Network was developed and a feature layer added scoring every within-category concept pair for visual and phonetic similarity. The system was evaluated with 21 children at a Sri Lankan special education school across 27 active days, producing 441 attempts and 247 mastery records. Tier 1 mastery was reached in 80.3 per cent of concept records and Tier 2 in 98.3 per cent, with a reliable within-concept improvement of 0.358 score points across repeated exposures (95 per cent CI 0.235 to 0.480; sign test $p < 0.001$). The principal finding concerns the similarity features: pairs children confused carried significantly higher visual similarity than pairs they did not ($z = 3.83$, $p = 0.00013$, $d = 0.62$), while phonetic similarity showed none ($p = 0.178$). The Graph Attention Network did not outperform the production heuristic or any baseline, and distractor ranking proved confounded by the exposure policy generating its own labels. The contribution is twofold: a deployed module validated with real autistic learners, and an honest negative result for graph attention alongside the finding that visual similarity carries confusion signal, resolving cold start for 96.3 per cent of untouched concepts.

Keywords: autism spectrum disorder, graph neural networks, adaptive learning, knowledge graph, special education technology

ACKNOWLEDGEMENTS

I wish to record my sincere thanks to my supervisor, Ms. Manori Gamage, and my co-supervisor, Ms. Suriya Kumari, whose guidance shaped both the direction of this research and the discipline with which it was evaluated. Their insistence that a claim be falsifiable before it is made has influenced this thesis more than any single technical decision recorded in it.

I am grateful to the principal, special education teachers, and staff of Thusitha Suwasetha Special Education School, Gampaha, who gave their time to the requirement gathering sessions, accommodated the system in live classroom practice, and provided the professional judgement against which the module's teacher-facing design was repeatedly corrected. The children who used the system, and the families who consented to their participation, made the evaluation reported in Chapter 6 possible.

My thanks go to my project group members, M. C. Gunawardena, R. S. Liluksha, and N. Samaranayake, for their collaboration on the shared platform within which this component operates, and for the integration work that allowed four independent modules to behave as one system.

Finally, I thank my family for their support and patience throughout this project.

TABLE OF CONTENTS

1 INTRODUCTION	1
1.1 Overall project context.....	1
1.2 Individual component context.....	2
1.3 Problem statement.....	3
1.4 Research gap	4
1.5 Aim and objectives.....	5
1.6 Research questions	6
1.7 Scope, limitations and assumptions	6
1.8 Individual contribution.....	7
1.9 Thesis structure	8
2 LITERATURE REVIEW AND RELATED WORK	9
2.1 Review method and scope.....	9
2.2 Behavioural and cognitive foundations.....	9
2.3 Machine learning approaches.....	10
2.4 Augmentative communication and user-centred design	12
2.5 Existing systems and products	13
2.6 Gap synthesis and conceptual framework.....	15
3 RESEARCH METHODOLOGY	17
3.1 Research design.....	17
3.2 Participants and study materials.....	18
3.3 Study procedure	19
3.4 Measures and evaluation metrics	20
3.5 Data analysis	21
3.6 Ethics, privacy, security and responsible AI.....	22

3.7 Reproducibility and research artefacts	23
4 PRODUCT AND SOLUTION DESIGN	25
4.1 Requirements.....	25
4.1.1 Users and stakeholders	25
4.1.2 Functional requirements.....	25
4.1.3 Non-functional requirements	27
4.2 Component architecture within the complete solution.....	28
4.3 Design alternatives and decisions	29
4.3.1 Training data source: graph versus event log.....	29
4.3.2 Unit of analysis: student versus concept pair	29
4.3.3 Model family: graph attention versus classical alternatives	30
4.3.4 Similarity features: computed versus learned	30
4.4 Detailed design.....	31
4.4.1 Tier progression and mastery	31
4.4.2 Adaptive difficulty and the relearn pathway	31
4.4.3 Graph Knowledge Base schema.....	32
4.4.4 Distractor selection chain	33
4.4.5 Concept similarity feature layer	34
4.5 Standards and professional practice	34
5 IMPLEMENTATION, INTEGRATION AND TESTING.....	36
5.1 Development environment and technology stack	36
5.2 Component implementation	37
5.2.1 Tier engine	37
5.2.2 Interaction logging	37
5.2.3 Graph synchronisation	37
5.2.4 Similarity feature pipeline.....	38

5.2.5 Teacher-facing analytics	38
5.3 Integration with the other components.....	38
5.4 Verification and software testing	39
5.5 Deployment and operation	41
6 RESULTS AND VALIDATION	42
6.1 Research results.....	42
6.1.1 Deployment scale and coverage.....	42
6.1.2 Tier mastery outcomes	43
6.1.3 Within-concept learning gain.....	46
6.1.4 Concept similarity and observed confusion	47
6.1.5 Engagement and scaffolding reliance	49
6.1.6 Category activity outcomes.....	50
6.2 Product and technical results.....	51
6.2.1 The exposure policy confound.....	51
6.2.2 Attempt outcome prediction.....	52
6.2.3 Mastery decision prediction	53
6.2.4 Summary of the graph intelligence result	54
6.3 User and stakeholder validation	54
6.4 Integration results.....	55
6.5 Summary of findings.....	55
7 DISCUSSION	57
7.1 Interpretation of findings.....	57
7.2 Comparison with literature, baselines and existing solutions	59
7.3 Research contribution.....	60
7.4 Product contribution.....	61
7.5 Limitations and threats to validity.....	62

7.6 Implications.....	63
8 COMMERCIALISATION AND SUSTAINABILITY	65
8.1 Value proposition and target users	65
8.2 Market and competitive position.....	65
8.3 Business and sustainability model	66
8.4 Scalability, intellectual property and compliance	67
8.5 Contribution to the Sustainable Development Goals	68
9 CONCLUSION AND FUTURE DIRECTIONS	69
9.1 Conclusion	69
9.2 Achievement of objectives.....	70
9.3 Individual research and product contributions	71
9.4 Future directions.....	71
REFERENCES.....	73
GLOSSARY.....	75
APPENDICES	76
Appendix A: Research instruments and extended results	76
A.1 Knowledge graph snapshot composition	76
A.2 Most strongly weighted observed confusions	76
A.3 Dataset versions used	77
A.4 Reproduction	77
Appendix B: Academic integrity and AI use disclosure	78

LIST OF FIGURES

Figure 6.1: Distribution of logged attempts across curriculum categories

Figure 6.2: Distribution of Tier 1 and Tier 2 mastery scores against the pass threshold

Figure 6.3: Concepts reaching the learnt state per participant

Figure 6.4: Direction of within-concept change across repeated exposures

Figure 6.5: Visual similarity for concept pairs observed as confused against pairs not confused

Figure 6.6: Audio replay and dwell time at Tier 1 compared with Tier 2

LIST OF TABLES

Table 1.1: Feature-level gaps identified in existing autism learning applications

Table 3.1: Concept categories and their composition

Table 4.1: Functional requirements for the Concept Learning Module

Table 4.2: Graph Knowledge Base relationship types

Table 5.1: Adaptive behaviour verification scenarios

Table 6.1: Per-participant mastery outcomes

Table 6.2: Similarity scores for confused against non-confused concept pairs

Table 6.3: Distractor ranking under leave-one-student-out cross-validation (confounded)

Table 6.4: Attempt outcome prediction under leave-one-student-out cross-validation

Table 9.1: Achievement of the specific objectives stated in Section 1.5

Table A.1: Knowledge graph node and relationship composition

Table A.2: Highest weighted confusion pairs by tier

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AAC	Augmentative and Alternative Communication
ABA	Applied Behavior Analysis
API	Application Programming Interface
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
BKT	Bayesian Knowledge Tracing
CI	Confidence Interval
CNN	Convolutional Neural Network
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GKB	Graph Knowledge Base
GNN	Graph Neural Network
IRT	Item Response Theory
JWT	JSON Web Token
LLM	Large Language Model
LOSO	Leave-One-Student-Out
MRR	Mean Reciprocal Rank
NIE	National Institute of Education
REST	Representational State Transfer
SDG	Sustainable Development Goal
SUS	System Usability Scale
TTS	Text to Speech

1 INTRODUCTION

1.1 Overall Project Context

Auriva is a tablet-first learning and support platform developed for children with Autism Spectrum Disorder (ASD) in Sri Lankan special education settings. The platform is built around a single organising decision: the teacher, not the parent and not the algorithm, runs the learning session. A special education teacher signs in, selects a child from their assigned list, and conducts the session on a shared tablet while the child interacts directly with the on-screen activities. The child never signs in, never navigates the application unaided, and never encounters an adult-facing control without passing a grown-up gate.

The platform delivers four learning modules, each developed as a separate research component by one member of the group. The Concept Learning Module, which is the subject of this thesis, teaches foundational English vocabulary. The Handwriting Module provides adaptive letter and word writing practice with motor-skill analysis. The Pronunciation Module scores spoken attempts using a phoneme-level goodness-of-pronunciation pipeline. The Dialogue Module teaches word use and sentence construction. The four modules share a common authentication layer, a common student and teacher data model, a common reporting surface in the Teacher Workspace, and a common design language.

The complete system is implemented as a React Native application supported by a Node.js REST API, three Python services, a PostgreSQL system of record, and a Neo4j graph database. The production deployment runs as a Docker Compose stack behind nginx, with the backend and the Python services reachable only on the internal network. The group project as a whole therefore constitutes a working, deployed software system rather than a prototype, and the evaluation reported in this thesis was conducted against that deployed system in a live classroom.

1.2 Individual Component Context

This thesis concerns the Concept Learning Module and the graph intelligence layer that supports it. The component covers the end-to-end path from a teacher selecting a concept category to the instructional decision about what that child sees next, and it comprises four distinct pieces of work.

The first is the tiered instructional architecture itself. Each vocabulary concept is taught through three tiers in a fixed sequence. Tier 1, Find it, establishes visual familiarisation: the child sees the concept image, hears the spoken word, watches a short demonstration, and then identifies the correct image among distractors. Tier 2, Name it, establishes single word recognition: the child hears the name and selects the written word, or drags the image onto the corresponding word. Tier 3, Watch and colour, introduces concept attributes through a short contextual video followed by an unscored colouring activity. A concept is recorded as learnt once Tier 1 and Tier 2 are both passed.

The second is the rule-based adaptive scaffolding engine. This enforces a mastery criterion of two out of three correct responses within a block before a tier may be passed, adjusts the number of distractors in response to consecutive correct or incorrect answers, implements a graduated three-level relearn pathway that varies the visual stimulus type rather than simply repeating the same item, and raises a teacher handover flag when a child reaches a persistent failure threshold rather than looping the child indefinitely.

The third is the Graph Knowledge Base (GKB), a Neo4j graph that models the learning domain as an interconnected structure of students, concepts, and categories, connected by typed, weighted relationships recording scores, engagement, and observed confusion. The GKB serves the adaptive engine at session time by supplying confusion-informed distractor candidates.

The fourth is the graph intelligence research strand: a Graph Attention Network (GAT) developed over the GKB as a shadow-mode prototype, evaluated against a complete suite of classical baselines, and accompanied by a concept similarity feature layer that scores every within-category concept pair for visual and phonetic

resemblance. This strand is where the research contribution of the component lies, and it is also where the most methodologically significant finding of this thesis emerged.

1.3 Problem Statement

Autistic children learning English as a second language in Sri Lankan special education settings are served by digital tools that are structurally unsuited to the task in four specific ways, each of which this component addresses.

Existing tools deliver content as isolated activities without a defined progression from recognition to naming to attribute understanding, and without a mastery requirement at one level before advancement to the next. A child may therefore pass through a vocabulary application without ever demonstrating durable competency at any stage, which undermines the reliability of the skill acquisition the tool claims to produce.

Error handling in these tools is superficial rather than genuinely adaptive. An application that renders an incorrect option unavailable after it is selected reduces a recognition task to elimination. An application that repeats the identical pronunciation after a failure changes neither the modality nor the stimulus. Neither constitutes the graduated, modality-varying remediation that the behavioural literature on autistic learning indicates is required.

No available tool models the relationships between the things being taught. A child who confuses apple with cherry is exhibiting a structured error, and that structure carries information about which other concepts will be difficult. A system that stores only per-item scores discards this information entirely. The consequence is most acute at cold start: when a child meets a concept for the first time, a per-item system has nothing to say about it, and the fallback is an arbitrary choice from the curriculum sequence.

Finally, these tools exclude the special education teacher from the instructional loop. Performance statistics, where they exist at all, are presented as raw numbers behind a subscription, with no interpretation, no concept-level breakdown, and no mechanism

by which a teacher can exercise professional judgement over the learning pathway the software has selected.

1.4 Research Gap

The feature-level audit conducted during the proposal phase of this project compared two widely used autism-focused learning applications, Otsimo and Speech Blub, against the requirements of a structured concept learning system. That audit is summarised in Table 1.1 and establishes six concrete gaps.

Table 1.1: Feature-level gaps identified in existing autism learning applications

Requirement	Otsimo	Speech Blub
Tiered mastery architecture	Absent; sequential activities only	Absent; competency-grouped only
Graduated relearn pathway	Absent; distractor removal only	Absent; identical repetition only
Sensory and typographic customisation	Absent	Absent
Teacher-mediated instructional role	Absent; parent-child model	Absent; parent-child model
Attribute-level concept classification	Absent	Absent
Relational model between concepts	Absent	Absent
Curriculum alignment to Sri Lankan syllabus	Absent; global content	Absent; global content
Offline operation	Absent; continuous connectivity required	Absent; continuous connectivity required

The cross-cutting absence is the final functional row of that table. Neither application, and none of the research systems reviewed in Chapter 2, models the conceptual relationships between learning items, the structural similarity between confusion patterns, or the cross-student evidence available when several learners exhibit the same difficulty cluster. This is the gap the present work addresses, and it is the gap that motivates a graph-based rather than a per-item representation.

A second, narrower gap concerns evidence. The literature contains several proposals for AI-driven adaptive learning for autistic children, and the strongest of them report high accuracy figures. Those figures are almost universally obtained on retrospective datasets or with very small numbers of neurotypical evaluators, rather than through

prospective deployment with autistic children in live classroom conditions. The present work reports the latter.

1.5 Aim and Objectives

The aim of this component is to design, implement, deploy, and evaluate an adaptive concept learning module that delivers curriculum-aligned English vocabulary instruction to autistic children up to Grade 3 level in Sri Lanka, in which adaptive decisions are informed by a graph representation of the learning domain and in which the special education teacher retains professional authority over the instructional pathway.

The specific objectives are the following.

1. To implement a three-tier concept learning structure covering visual familiarisation, single word recognition, and attribute introduction, sequenced across curriculum-aligned concept categories derived from the Sri Lankan National Institute of Education Grade 1 and Grade 2 English syllabus.
2. To design and populate a Graph Knowledge Base comprising a static curriculum layer and a dynamic interaction layer updated from live session logs.
3. To develop a rule-based adaptive scaffolding engine that applies an Applied Behaviour Analysis mastery criterion, adjusts activity difficulty in response to performance, and implements a graduated relearn pathway that varies stimulus modality.
4. To construct a concept similarity feature layer scoring every within-category concept pair for visual and phonetic resemblance, and to test empirically whether either similarity dimension predicts the confusions children actually make.
5. To develop and evaluate a Graph Attention Network over the Graph Knowledge Base against a complete suite of classical baselines under leave-one-student-out cross-validation, and to determine whether graph attention improves on the deployed heuristic at the available data scale.
6. To provide teacher-facing interpretive feedback and per-child progress reporting that supports, rather than replaces, professional instructional judgement.

7. To deploy the module in a live special education setting and evaluate it with autistic children under conditions of informed consent and ethical oversight.

1.6 Research Questions

The research strand of this component is organised around three questions.

RQ1. Does a structured, tier-gated, curriculum-aligned concept learning module produce measurable within-concept learning gains when deployed with autistic children in a live special education classroom?

RQ2. Do computed concept similarity features, specifically visual and phonetic similarity, predict the concept confusions that autistic children actually exhibit during learning?

RQ3. Does a Graph Attention Network over the Graph Knowledge Base outperform the deployed rule-based heuristic and classical baselines on distractor selection and attempt-outcome prediction at the data scale achievable in this deployment?

RQ1 and RQ2 are answered affirmatively in Chapter 6. RQ3 is answered negatively, and the analysis of why it is answered negatively, reported in Section 6.2 and discussed in Section 7.1, constitutes a substantive methodological contribution of this work.

1.7 Scope, Limitations and Assumptions

The scope of this component is bounded in the following ways. The vocabulary domain is restricted to foundational English concepts up to Grade 3 level across nine curriculum categories. The system is designed for teacher-mediated use on a shared tablet and is not intended for independent child use or for home deployment without a supervising adult. Tier 3 colouring activities are unscored by design, because a colouring task has no single correct outcome and scoring it would convert an expressive activity into an assessment.

Three limitations are inherent to the evaluation and are examined in detail in Section 7.5. The participant group comprises 21 registered children drawn from a single special education school, which constrains generalisation. Interaction data is

unevenly distributed across curriculum categories, with the fruits category accounting for a substantial majority of recorded attempts. The deployment period of 27 active days does not permit assessment of long-term retention.

Three assumptions underpin the work. It is assumed that the teacher present during each session is the appropriate authority on whether the child is engaged and whether a session should continue, and the system is designed to defer to that judgement. It is assumed that the curriculum content provided by the National Institute of Education is the correct instructional target, as confirmed by the client requirements document. It is assumed that behavioural error patterns captured through interaction logging are a valid signal of conceptual difficulty, which is the methodological premise on which the entire graph layer rests, and which is defensible precisely because verbal self-report is frequently unavailable in this population.

1.8 Individual Contribution

The contributions of this component are the following.

- A deployed three-tier concept learning module, aligned to the Sri Lankan national curriculum, operating under teacher mediation, and validated with 21 registered autistic learners over 27 active days of classroom use.
- A rule-based adaptive scaffolding engine implementing an Applied Behaviour Analysis mastery criterion with a graduated, modality-varying relearn pathway and a teacher handover mechanism that bounds the number of consecutive failures a child may experience.
- A Graph Knowledge Base modelling students, concepts, and categories with typed, weighted relationships, extended during this work with an append-only attempt history that renders the graph temporal and therefore reproducible for research use.
- A concept similarity feature layer, and the empirical finding that visual similarity predicts observed confusion with a medium effect size while phonetic similarity does not. This finding is directly actionable: it identifies which feature should drive distractor selection, and it resolves the cold-start problem for 96.3 per cent of concepts no child has yet attempted.

- A complete, reproducible evaluation harness implementing leave-one-student-out cross-validation against five classical baselines, and a negative result for Graph Attention Networks at this data scale, reported with the exposure-policy confound that explains it.
- The identification and documentation of that confound, namely that the system's own distractor fallback policy constructs the labels against which distractor quality would be measured, rendering offline ranking claims unfalsifiable without instrumented option-set logging.

1.9 Thesis Structure

Chapter 2 reviews the behavioural and cognitive foundations of autism learning technology, surveys machine learning approaches in this domain, audits existing commercial products, and synthesises the gap addressed by this work. Chapter 3 sets out the research methodology, including the mixed-methods design, the participants, the evaluation protocol, and the ethical framework governing work with a vulnerable population. Chapter 4 presents the design of the component, covering requirements, architecture, the design alternatives considered, and the detailed design of the tier engine and graph layer. Chapter 5 documents implementation, integration with the other three components, verification testing, and deployment. Chapter 6 reports results, covering learning outcomes, the similarity analysis, the graph attention evaluation, and integration results. Chapter 7 interprets those results, compares them with the literature, states the research and product contributions, and examines limitations and threats to validity. Chapter 8 addresses commercialisation and sustainability. Chapter 9 concludes and sets out future directions.

2 LITERATURE REVIEW AND RELATED WORK

2.1 Review Method and Scope

This review covers two distinct bodies of evidence. The first is the scholarly literature on technology-supported learning for children with ASD, with emphasis on behavioural foundations, machine learning approaches, and augmentative communication. The second is the commercial product landscape, audited at feature level, because the practical alternatives available to a Sri Lankan special education teacher are products rather than papers.

Scholarly sources were identified through IEEE Xplore, ACM Digital Library, and Google Scholar using combinations of the terms autism, ASD, adaptive learning, vocabulary acquisition, graph neural network, and special education technology, restricted to peer-reviewed conference and journal publications. Sources were retained where they reported either an implemented system evaluated with users or a methodological contribution bearing on adaptive instruction for this population. Product sources were evaluated by direct hands-on feature audit of the installed applications rather than by reference to marketing material, because published feature lists for consumer applications are unreliable.

The review is organised into four substantive sections covering behavioural foundations, machine learning approaches, augmentative communication research, and user-centred design, followed by the product audit and the gap synthesis that closes the chapter.

2.2 Behavioral and Cognitive Foundations

Hariharan et al. [4] provide a comparative study of 196 autistic and 196 non-autistic adults using the Camouflaging Autistic Traits Questionnaire, documenting significant behavioural differences across masking, compensation, and assimilation dimensions. Two of their findings bear directly on instructional design. Autistic individuals in their sample relied heavily on structured external supports and showed low comfort in unstructured contexts, and they demonstrated a strong tendency to learn through observation and repetition rather than through spontaneous social

exchange. These are not incidental preferences. They indicate that structured, predictable sequences, repeated audio-visual exposure, and guided feedback are necessities for this population rather than usability refinements, and the three-tier architecture described in Chapter 4 is a direct consequence.

Wang and Wang [5] review the etiology, diagnosis, and therapy of ASD, and identify Applied Behaviour Analysis (ABA) as among the most extensively evidenced therapeutic frameworks. ABA-based interventions, including discrete trial training, pivotal response treatment, and early intensive behavioural intervention, share a commitment to decomposing target skills into small sequentially mastered steps, using reinforcement to consolidate learning, and maintaining high mastery thresholds to ensure genuine retention rather than transient performance.

The mastery threshold question is treated specifically by Granpeesheh et al. [6] and by Fuller and Fienup [7]. Fuller and Fienup examine the effect of mastery criterion level on response maintenance and report that higher criteria produce more durable retention. The present system adopts a criterion of two correct responses out of three within a block, applied consistently at each tier. This is a deliberate calibration rather than a direct adoption of the strictest criterion in the literature: a very high threshold applied to a child who is already struggling produces extended failure sequences, and for an autistic learner a sustained failure sequence carries a clinical cost in frustration and task refusal that the behavioural literature on retention does not price.

This tension, between the retention benefit of a strict criterion and the frustration cost of an unreachable one, is resolved in the present design by pairing a moderate criterion with a graduated relearn pathway and a bounded failure count terminating in teacher handover. The system therefore maintains a mastery requirement while guaranteeing that no child encounters an unbounded sequence of failures.

2.3 Machine Learning Approaches

Vijayan et al. [8] propose a three-layer cognitive computing architecture for an intelligent learning assistant for children with ASD, integrating an application layer with a chatbot and visual aids, a data engineering layer processing facial images and brain images, and an AI layer employing R-CNN for facial pattern recognition, deep

CNN for brain image classification, and reinforcement learning with deep Q-learning for adaptive courseware delivery. The work demonstrates that AI-driven adaptive learning for ASD populations is technically feasible and theoretically well motivated. Its limitation is that the framework remains largely architectural, with limited evidence of empirical validation with autistic children in real classroom settings. The present system deliberately restricts its scope to what can be implemented and validated within the Sri Lankan special education context, and reports that validation in Chapter 6.

Hayat et al. [9] apply machine learning and deep learning to personalised teaching strategy suggestion for toddlers with ASD. Using a combined dataset of 2,242 instances with feature selection via ANOVA F-values, they identify executive function, language and communication, and eye contact as the strongest predictors of teaching strategy suitability. Their models achieve accuracy rates of up to 100 per cent on the combined dataset. That figure warrants scrutiny rather than emulation: a model reporting perfect accuracy on a retrospective dataset is more likely to indicate label leakage or an unrepresentative split than genuine predictive power, and the authors themselves acknowledge that their models are trained on existing datasets rather than prospectively validated in live teaching environments, and that ground-truth labels for teaching strategy appropriateness remain dependent on clinical judgement.

Ganju et al. [10] present CogniQuest, a CNN-based iOS application for neurodivergent children aged six to twelve using handwriting recognition to assess cognitive skills. Their architecture achieves a training accuracy of 98.85 per cent and a validation accuracy of 97.70 per cent on the Kaggle A-Z handwritten alphabet dataset. The technical result is strong, but their user evaluation is limited to three neurotypical children, which raises a substantial generalisation question for autistic populations whose motor profiles, cognitive processing styles, and engagement patterns differ materially from neurotypical peers. The contrast with the present work is instructive: this thesis reports lower headline numbers obtained from 21 autistic children in a live classroom, which is the harder and more informative evidence.

Graph neural networks have not, to the knowledge of this review, been applied to concept confusion modelling in autism education specifically. The relevant methodological grounding is therefore drawn from the general graph learning literature. Graph Attention Networks, introduced by Velickovic et al., compute each node representation as an attention-weighted sum of its neighbours, learning how much each neighbour contributes rather than averaging uniformly. This property is directly relevant here, because the production system already encodes a hand-set attention weight: personal confusion evidence is weighted three times pooled cohort evidence, a ratio chosen by judgement rather than learned from data. The motivation for attention over simpler aggregation in this domain is precisely that this ratio should be learned per concept pair rather than applied uniformly.

Two cautionary citations govern the interpretation of attention weights in this work. Jain and Wallace argue in *Attention is not Explanation* that attention distributions do not reliably identify the inputs responsible for a prediction, and Wiegrefe and Pinter reply in *Attention is not not Explanation* that the conclusion depends on what one requires of an explanation. The position adopted in this thesis is that attention weights may serve as a plausible-hypothesis generator for teachers, always subject to professional override, and never as a claim about the child's cognition.

2.4 Augmentative Communication and User-Centered Design

Kumar et al. [11] propose an AAC-based big data analytics framework for children with neural disorders, incorporating eye-tracking via a Tobii X2-60 device to enable gaze-based communication support. Their CNN model achieves 98.63 per cent accuracy in predicting communicative intent from gaze data. The pipeline, from gaze collection through preprocessing and feature extraction to text-to-speech output, is technically sophisticated. It presupposes access to eye-tracking hardware that is unavailable in most Sri Lankan special education settings, which is a decisive constraint for the present work: a system intended for deployment in government special education units must operate on commodity tablets.

Hiryanto et al. [12] describe VICARA 3.0, a redesigned Bahasa Indonesia AAC mobile application for non-verbal autistic children. Their design thinking approach

produces a system evaluated against nine ASD-specific UX guidelines covering engaging, predictable, structured, interactive, generalizable, customizable, sense aware, attention retaining, and frustration free properties, confirmed compliant by a focus group including an ICT representative, an ASEAN Autism Studies representative, and a speech-language pathologist. Notable design features include Fitzgerald colour-coded picture cards, emotion-representing mascots, selectable visual themes, a PIN-protected parental mode, and text-to-speech integration. Their commitment to structured, predictable navigation and avatar-based guidance informs the present system's interface design directly, though VICARA focuses exclusively on AAC communication rather than curriculum-aligned language learning.

Akbar et al. [13] apply User-Centred Design methodology to a scheduling activity application for autistic children aged seven to twelve. Their dual-interface Android application, comprising a parent-facing content creation interface and a child-facing activity-following interface, is evaluated using the QUIM usability model across ten factors in two iterative design cycles, achieving final scores of 90.5 per cent for parents and 89.9 per cent for children. Their most transferable finding is that splitting the parent input process across multiple views improved the effectiveness dimension from 84 per cent to 90 per cent, which indicates that task decomposition matters as much in the adult interface as it does in the child interface. The present system adopts the same dual-perspective architecture, maintaining a strictly separated Teacher Workspace and Student Workspace, with the teacher interface designed to support professional decision-making rather than to mirror the child's experience.

2.5 Existing Systems and Products

Otsimo is a commercially available mobile learning application for autistic children. The feature audit conducted for this research found genuine competence across several dimensions of instructional design: a structured activity selection interface, a parent-gated settings panel secured through numeric adult authentication, child profiles accessible only within the parent-controlled environment, a premium-tiered multi-user architecture, and an onboarding questionnaire spanning cognitive and social developmental indicators that generates a personalised curriculum.

Reinforcement design is multimodal, comprising confetti animations, per-step star allocation, achievement badges, and positively valenced verbal feedback. Instructional content spans vocabulary matching, cloze exercises, listening comprehension, bidirectional picture-word correspondence, and drag-and-drop interaction.

Against these strengths, the audit surfaced limitations with direct relevance to the present work. Navigation lacks category-based organisation, confining the learner to sequential progression through individual activities. No provision exists for typographic accessibility: adjustable text size, font selection, and high contrast modes are uniformly absent. Auditory scaffolding options including slow speech mode and variable speech rate are unavailable. The content architecture is closed, precluding custom vocabulary addition, learner-specific image upload, or editable word sets, which means a teacher cannot align the tool to a national curriculum. The age verification mechanism does not distinguish verified guardians from other adults. Continuous internet connectivity is required. Most significantly for the present context, Otsimo is not architecturally conceived as a teacher-mediated instrument and provides no mechanism by which a special education teacher can exercise professional oversight.

Speech Blub is oriented towards speech and language development. Its evaluated strengths lie in interface personalisation and auditory instruction: the application generates a bespoke child avatar from a submitted photograph which then appears throughout the application as a participatory model, a design approach with theoretical grounding in social learning. Pronunciation modelling is supported through slowed speech repetition, audio instructions are delivered via the avatar at each stage, and the interface provides visual cue highlighting and generously proportioned interactive targets. A parent gate is implemented using date-of-birth verification with session timeout, and an initial assessment questionnaire spanning hearing, pragmatics, play, comprehension, and expressive language produces a developmental profile benchmarked against age-typical norms.

The functional gaps are nonetheless substantial for structured educational use. Text-to-speech output, symbol-based communication support, word and sentence

construction tools, and visual vocabulary card systems are absent. Category-based navigation is only partially realised, organised by broad competency domain rather than discrete sequentially structured content categories. A notable safety concern is the absence of authenticated return logic on the back button, permitting unrestricted child-initiated navigation without adult re-verification. Typographic and visual customisation is entirely unavailable. Performance data collection is limited: neither accuracy rates nor time-on-task metrics are recorded at activity level. Educator access to learner progress is constrained by the absence of a dedicated reporting interface.

2.6 Gap Synthesis and Conceptual Framework

Across the scholarly literature and the commercial product landscape, six interconnected gaps emerge, and the proposed system addresses all six within a single integrated platform.

First, neither audited application provides a structured, tiered concept learning architecture with a defined progression from visual familiarisation through word recognition to attribute understanding, and neither requires demonstrated competency at one level before advancement. Second, error handling in both remains superficial and does not constitute genuine adaptive scaffolding, with no graduated relearn pathway that varies stimulus modality across attempts. Third, sensory customisation options are inadequate, despite the established sensitivity of autistic children to visual and typographic presentation. Fourth, both operate exclusively within a parent-child model and provide no role for the special education teacher. Fifth, neither extends concept learning beyond surface recognition to defining attributes. Sixth, neither conducts structured post-concept evaluation assessing retention and generalisation across multiple previously learned concepts.

The cross-cutting gap, and the one that motivates the research strand of this component, is the absence of any relational or graph-based adaptive mechanism in any reviewed system. Neither Otsimo nor Speech Blub models the conceptual relationships between learning items, the structural similarity between confusion patterns, or the cross-student evidence available when multiple learners exhibit the

same difficulty cluster. The conceptual framework adopted in this work follows from that absence directly: if confusion between concepts is a structured behavioural signal, then the appropriate representation is a graph in which concepts are nodes and observed confusions are weighted edges, and the appropriate question is whether that structure can be exploited to make better instructional decisions than per-item scoring allows.

Chapter 6 reports that it can, but not in the way originally hypothesised. The graph structure proves valuable through its feature layer rather than through learned graph attention, and the reasons for that distinction form the substance of Chapter 7.

3 RESEARCH METHODOLOGY

3.1 Research Design

This study adopts a mixed-methods research design, combining quantitative system and learning evaluation with qualitative user-centred assessment. The design is necessary because the research problem requires evidence along two distinct dimensions that no single methodology addresses.

The first dimension is whether the system performs as intended. The Concept Learning Module implements tier-based progression, adaptive difficulty adjustment, graduated relearn pathways, distractor confusion tracking, and graph-informed decision support. Each produces measurable outputs verifiable through functional testing and system log analysis, and each learning interaction produces a scored record from which learning outcomes can be computed.

The second dimension is whether the system is usable and appropriate for its target users. The end users are young autistic children, and the proxy users are special education teachers who oversee every session. Whether the interface accommodates sensory sensitivities, whether personalisation makes a practical difference, whether teachers find the generated feedback actionable, and whether the system supports rather than undermines professional judgement are questions that system logs cannot answer.

A purely quantitative approach would confirm that the system functions correctly while leaving its real-world suitability for a vulnerable population unvalidated. A purely qualitative approach would capture user experience without verifying that the underlying learning logic operates to specification. The mixed-methods approach produces two complementary bodies of evidence.

Within the quantitative strand, the study is observational rather than experimental. No randomised controlled trial was conducted, and none could have been: a participant group of this size cannot support a between-subjects randomised design with adequate statistical power, and withholding an instructional tool from a control group of children with identified learning needs raises an ethical objection that the

research question does not justify. The quantitative analysis is therefore within-subject, comparing each child's performance across repeated exposures to the same concept, and the inferential claims in Chapter 6 are restricted to what that design supports. This limitation is examined further in Section 7.5.

3.2 Participants and Study Materials

Participants were children registered in the deployed system at Thusitha Suwasetha Special Education School, Gampaha, the client organisation that participated in requirement gathering for this project. Twenty-one child records exist in the production knowledge graph at the time of the evaluation snapshot. Two of these are non-participant accounts: one account created for system testing and one account carrying no participant identity. These were excluded from all analyses reported in Chapter 6, leaving 19 participant records with tier-level learning data.

Participation was mediated throughout by the child's assigned special education teacher, who initiated every session, remained present for its duration, and retained authority to terminate it. Seventeen teachers are registered in the wider system; the concept module sessions reported here were conducted by the subset of those teachers assigned to the participating children.

The study materials are the concept assets of the Concept Learning Module itself. The curriculum comprises 96 concepts organised into nine categories, derived from the Grade 1 and Grade 2 English Listening and Speaking Teacher's Guides published by the National Institute of Education of Sri Lanka, as specified in the client requirements document. Each concept carries a real-life photograph, an illustrated cartoon variant, a short contextual animation, recorded narration audio of the concept word, and colour and shape attribute metadata. Category composition is given in Table 3.1.

Table 3.1: Concept categories and their composition

Category	Concepts
animals	21
household	14

Category	Concepts
classroom	11
fruits	11
colors	10
numbers	10
professionals	8
house	4
shapes	4
Total (categorised)	93

Three further concepts in a family category were created during the deployment period at teacher request and are present in the graph without a category relationship. The total concept count in the knowledge graph is therefore 96.

3.3 Study Procedure

The procedure followed the ordinary operation of the deployed system rather than an imposed experimental protocol, because the research question concerns how the module behaves in real classroom use.

On a child's first session, the system conducts a one-time preference assessment capturing avatar choice, visual theme, and font style and size. These preferences are stored against the student record and applied consistently in all subsequent sessions, so that the interface remains predictable and sensorially comfortable. No learning data is collected during this assessment.

In each learning session, the teacher selects the child, selects a concept category, and selects a concept. The system resumes at the tier the child reached previously. Within Tier 1, the child is shown the concept image, hears the spoken word, may replay the audio at will, watches a demonstration, and then completes identification trials organised in blocks. Within each block the system presents the target among distractors drawn from the distractor selection chain described in Section 4.4. Mastery is assessed against the two-of-three criterion within a block. On failure to meet the criterion, the graduated relearn pathway is triggered: Level 1 switches the

visual stimulus from photograph to illustrated cartoon, Level 2 introduces a short contextual animation before retrying, and Level 3 returns the child to the full familiarisation screen with animation-on-tap interaction. Tier 2 follows the same block and mastery structure for word recognition. Tier 3 presents a contextual video followed by an unscored colouring activity.

Every interaction event is logged with its timestamp, the active tier, the concept, the response, the selected distractor where the response was incorrect, the active difficulty level, and the elapsed time. Logs are written to PostgreSQL as the system of record and projected into the Neo4j Graph Knowledge Base, where each attempt is additionally persisted as a first-class Attempt node, so that the graph retains history rather than only current state.

Sessions were conducted over 27 active days between 2 May 2026 and 9 October 2026. No session length or session frequency was imposed by the research protocol; both were determined by the teacher according to the child's engagement and the school timetable.

3.4 Measures and Evaluation Metrics

The following measures are reported in Chapter 6.

Tier mastery score is the proportion of correct responses within the assessed block, recorded on a scale from 0 to 1 and compared against the pass threshold of two thirds. Pass rate is the proportion of tier records meeting that threshold. Concepts learnt is the count of concepts for which a child passed both Tier 1 and Tier 2.

Within-concept learning gain is the difference between the first and last recorded score for a given child, concept, and tier where more than one attempt exists. This is the primary learning outcome measure, and it is a within-subject measure by construction, which is what makes it defensible at this sample size.

Engagement measures comprise audio replay and tap counts and dwell time per concept, recorded separately for Tier 1 and Tier 2. These are behavioural proxies for attention and scaffolding reliance rather than direct measures of either.

Confusion is recorded as a weighted directed edge between two concepts whenever a child selects one when the other was correct, accumulated separately for Tier 1 identification, Tier 2 name recognition, activity-level errors, and image-format-specific errors.

For the graph intelligence strand, distractor ranking is evaluated using Hits@1, Hits@2, and Mean Reciprocal Rank. Attempt-outcome prediction is evaluated using accuracy, log loss, Brier score, and ROC-AUC. Bare accuracy is not reported as a primary metric for outcome prediction, because the class distribution is heavily imbalanced and a majority-class predictor attains a misleadingly high score.

For the similarity analysis, visual and phonetic similarity scores for each within-category concept pair are compared between pairs that were observed to be confused and pairs that were not, using the Mann-Whitney U test with a normal approximation, reported alongside the rank-biserial correlation and Cohen's d as effect size measures.

3.5 Data Analysis

Learning outcome analysis treats the concept-tier record as the unit of analysis. Descriptive statistics are reported for score distributions, pass rates, and per-participant concept counts. Within-concept gain is tested using a two-sided sign test on the direction of change, which makes no distributional assumption, and reported alongside the mean gain with a t-based 95 per cent confidence interval. Tier 1 and Tier 2 scores are compared using a paired comparison restricted to the student-concept pairs present at both tiers.

The similarity analysis compares two groups of concept pairs. The unit of analysis is the undirected concept pair rather than the child, which is a deliberate methodological choice: it pools evidence across all participants and so is not limited by the number of children, and it is the framing under which the available data supports an inferential claim. Directed similarity edges are collapsed to undirected pairs before analysis to avoid counting each pair twice.

Graph attention evaluation uses leave-one-student-out cross-validation, in which the model is trained on all students but one and evaluated on the held-out student, repeated for every student. Cohort statistics are computed from other students only, and personal statistics from the held-out child's own earlier observations only, so that no information from the evaluation fold leaks into training. Each configuration is run with ten random seeds and results are reported as means, because single-run figures at this data scale are not meaningful. Five classical baselines are implemented for comparison: the production heuristic, a global confusion popularity prior, a Rasch one-parameter item response model, Bayesian Knowledge Tracing, and concept difficulty, alongside Graph Convolutional Network and GraphSAGE ablations that isolate the contribution of attention specifically.

All analysis code operates on an exported, versioned snapshot of the production data rather than on the live database, so that every figure reported in Chapter 6 is reproducible from a fixed input.

3.6 Ethics, Privacy, Security and Responsible AI

This study involves minors with a developmental disability, and the ethical requirements are correspondingly strict.

Written parental or guardian consent was obtained for child participants, and assent was sought from the child where their developmental level permitted meaningful assent. Informed written consent was obtained from participating teachers. Participation was voluntary at every stage and withdrawal was available without consequence. Any session in which a child displayed distress, frustration beyond what is typical for the task, or unwillingness to continue was terminated immediately by the supervising teacher, regardless of whether the activity was complete. This authority rested with the teacher and not with the researcher or the system.

No personally identifiable information about any child is transmitted to external services. Prompts sent to the language model used for teacher-facing summaries carry anonymised session data excluding names, identification numbers, and any other identifying detail. All collected data is stored in access-controlled

infrastructure, available only to the research team, and is not disclosed to third parties.

Three constraints govern the responsible use of the machine learning components, and each is enforced in the deployed system rather than stated as an aspiration. The model never gates tier progression; progression remains under teacher and threshold control, because a false negative would lock a child out of content they are ready for. The model never overrides sensory settings, which are a clinical decision and not an optimisation target. The model operates within a hard timeout with fallback to the deployed heuristic, so that a slow or failed inference degrades invisibly to existing behaviour rather than interrupting a child's session.

A further constraint governs the teacher-facing language model output: it is instructed to describe and never to diagnose. Generated summaries report what the session data shows in plain language and do not offer clinical interpretation, developmental comparison, or deficit framing.

Two ethical limitations are recorded honestly. The consent obtained for classroom deployment covers use of the system and collection of learning data; the question of whether it extends to secondary use of that data for model development was identified during this work as requiring explicit confirmation rather than assumption, and the graph attention evaluation reported in Section 6.2 was conducted on pseudonymised data for that reason. Second, the retention and deletion policy for models trained on data from a child who subsequently withdraws is a question this work raises but does not fully resolve, and it is listed in Section 9.4 as required future work.

3.7 Reproducibility and Research Artefacts

The evaluation reported in this thesis is reproducible from fixed artefacts. The dataset builder replays the interaction event log into dated graph snapshots and writes a versioned dataset with a content hash for each table, so that a later run producing different numbers can be distinguished from a later run on different data. Four dataset versions were produced during the research period, and the version used for each result is stated where that result is reported.

The evaluation harness implements the leave-one-student-out protocol, all five classical baselines, and the three graph architectures behind a single command, with fixed random seeds. The similarity feature layer and the descriptive analyses reported in Section 6.1 operate on an exported snapshot of the knowledge graph taken on 11 October 2026, comprising 570 nodes and 2,663 relationships.

The research code is maintained separately from the serving code. The inference service carries only the dependencies required to serve a frozen model artefact, and the training dependencies, which are substantially larger, are confined to an offline requirements specification that the production container does not install.

4 PRODUCT AND SOLUTION DESIGN

4.1 Requirements

4.1.1 Users and Stakeholders

Three roles interact with the system. The principal acts as system administrator, creating and managing teacher and student accounts and assigning students to teachers. The principal does not interact with learning content. Special education teachers are the primary operational users and are classified as proxy users: they navigate and manage the system on behalf of the child during learning sessions, initiate sessions, select modules, monitor responses, and review performance feedback. Autistic children are the end users and direct beneficiaries. They do not navigate the system independently, but they are the ones who physically interact with the learning activities, and all content, interface design, and adaptive mechanisms are designed around their needs.

4.1.2 Functional Requirements

The functional requirements specific to this component are listed in Table 4.1. Requirements relating to authentication, account management, and the shared teacher dashboard are the responsibility of the wider group project and are not restated here.

Table 4.1: Functional requirements for the Concept Learning Module

ID	Requirement
FR1	The system shall organise concepts into categories sequenced by increasing cognitive complexity, derived from the NIE Grade 1 and Grade 2 English curriculum.
FR2	The system shall pre-classify every concept by attribute stability as fixed, dominant, or variable, and use this classification to determine which Tier 3 activities are presented, adapted, or omitted.
FR3	The system shall deliver Tier 1 by displaying a concept image that pronounces the concept word on tap for up to five repetitions, defaulting to a real-life photograph.
FR4	The system shall assess Tier 1 mastery through an audio-visual identification activity structured in blocks, applying a two-of-three accuracy criterion, with image positions randomised between attempts.

ID	Requirement
FR5	The system shall dynamically adjust distractor count within Tier 1, increasing on consecutive correct answers and reducing on consecutive incorrect answers.
FR6	The system shall implement graduated, non-punitive error handling within each block, progressing from a subtle visual indication, to distractor removal, to correct-answer highlighting with a hint.
FR7	The system shall trigger a three-level relearn pathway on failure to meet the mastery threshold, where Level 1 switches from photograph to illustrated cartoon, Level 2 introduces a contextual animation, and Level 3 returns the child to full familiarization.
FR8	The system shall record the image format that proves most effective for each child in Tier 1 and use that format as the default visual stimulus in Tier 2 for the same concept.
FR9	The system shall deliver Tier 2 through drag-and-drop of the image onto the corresponding word followed by word-to-image matching, applying the same mastery criterion and block structure.
FR10	The system shall track word-visual confusion during Tier 2 and, on detecting visually similar word confusion, switch to more distinct fonts, increase spacing between word options, and highlight the initial letter of each word.
FR11	The system shall deliver Tier 3 by identifying the concept category, displaying a short animation demonstrating real-world function, and introducing colour and shape attributes through interactive activities.
FR12	The system shall flag a concept for teacher intervention if a child reaches persistent failure thresholds at any tier after exhausting all relearn levels, and shall not loop the child indefinitely through repeated attempts.
FR13	The system shall allow the teacher to decide whether to return the child to a previous tier, reattempt in a future session, or temporarily move forward, maintaining teacher authority over algorithmic recommendation.
FR14	The system shall log audio replay frequency, distractor selection patterns, relearn level reached, relearn cycle count, and mastery attempt count per tier for every concept worked on by every child.
FR15	The system shall maintain a Graph Knowledge Base comprising a static curriculum layer and a dynamic interaction layer, updating edge weights in the dynamic layer following each interaction block.
FR16	The system shall persist each scored attempt as an append-only record in the knowledge graph, so that interaction history is retained rather than overwritten by

ID	Requirement
	current state.
FR17	The system shall compute visual and phonetic similarity scores for every within-category concept pair and store them as graph relationships available to the distractor selection chain.
FR18	The system shall provide inline language-model-generated interpretive summaries beneath each concept view in the teacher interface, written in plain, non-technical, non-diagnostic language.
FR19	The system shall provide a per-child statistics dashboard displaying category-level progress, concept-level tier completion status, cumulative relearn cycle counts, and evaluation scores.
FR20	The system shall enable the teacher to generate a formal report for a selected child over a specified date range.

4.1.3 Non-Functional Requirements

Usability requires calming, sensory-friendly colour palettes across all interface elements, a minimal uncluttered interface with no complex navigation structures, clearly visible interactive elements adequately sized for touch, and freedom from irrelevant animated or decorative distraction. Accessibility requires support for multiple image formats for the same concept to accommodate individual differences in visual processing, and clear, appropriately paced audio delivered in a consistent voice, available as voluntary replay at all times.

Performance requires that learning activities render without latency that could interrupt engagement or cause confusion during timed interaction sequences, and that graph inference and language model generation never block a child-facing interaction. Data integrity requires that all session data persist reliably and that the difficulty level active at the time of every score is recorded alongside it, so that performance data remains interpretable in its original context.

Availability carries a specific meaning in this system: a child's learning session must not be interrupted by background processes such as language model feedback

generation, report compilation, or data synchronisation. Teacher-facing analytics are isolated from child-facing learning activity execution.

4.2 Component Architecture Within the Complete Solution

The Concept Learning Module spans four tiers of the platform architecture, and its design is shaped throughout by one constraint: nothing in the intelligence layer may be allowed to delay or interrupt what the child sees.

At the presentation tier, the module contributes the category selection screen, the concept selection screen, the three tier activity screens, and the category game screens, all implemented within the shared React Native application. The child-facing screens lock to landscape orientation and apply the student's stored sensory preferences on session start.

At the application tier, the Node.js backend hosts the concept service and the activity service. The concept service owns tier progression, mastery evaluation, relearn pathway state, and the distractor selection chain. The activity service owns the category games and the difficulty ladder. Both write to PostgreSQL synchronously and to the graph service asynchronously.

At the intelligence tier, a separate FastAPI service owns the Graph Knowledge Base. It exposes graph read and write endpoints over Neo4j and is called by the Node.js backend. Every call to it is made in a fire-and-forget pattern for writes and under a hard 400 millisecond timeout for reads, with a defined fallback for every read. This is the architectural expression of the constraint stated above: if the graph service is slow, unreachable, or returns nothing useful, the child's session proceeds on the fallback path and no error is surfaced to the child.

At the data tier, PostgreSQL is the system of record and holds the authoritative interaction log, progress records, and report artefacts. Neo4j holds the graph projection. This separation is deliberate and its rationale is given in Section 4.3.

Integration with the other three components of the group project occurs at two surfaces. The shared student and teacher model and the authentication layer are common. The Teacher Workspace presents concept progress as one tab among four,

using the same card and report layout as the handwriting, pronunciation, and dialogue modules, so that a teacher encounters one consistent reporting surface rather than four.

4.3 Design Alternatives and Decisions

Four design decisions in this component involved substantive alternatives, and each is recorded here with its rationale.

4.3.1 Training Data Source: Graph Versus Event Log

The initial design intention was to train the graph model directly on the knowledge graph. Analysis of the graph write path established that this was not possible. Every score and engagement edge was written with an update-on-match semantic, meaning each new attempt overwrote the previous value. The graph held current state, not history: 60 score edges represented 60 final states, not 60 observations. A model trained on it would see a single snapshot with no temporal dimension, and could not learn that a child confused two concepts in May but not in August.

Three alternatives were considered. The first was to use the PostgreSQL interaction log as the training corpus and replay it into dated snapshots, requiring no schema change. The second was to add attempt nodes to the graph so that the graph itself became temporal, which is cleaner but imposes additional write load on a constrained database tier. The third was nightly export to columnar files for reproducibility only.

The decision was to implement the first immediately, because it unblocked all downstream work without a schema change, and subsequently to implement the second as well. The graph snapshot analysed in Chapter 6 contains 444 append-only attempt records, confirming that the graph is now temporal. The third was implemented as the versioned dataset builder described in Section 3.7.

4.3.2 Unit of Analysis: Student Versus Concept Pair

The second decision determined whether the research strand was viable at all. Framing the prediction task per student yields 19 training examples, which is hopeless. Framing it per student-concept pair yields 66 examples with a single

negative case, which is untrainable: a classifier predicting pass for every row scores 89 per cent accuracy while learning nothing. Framing it per attempt yields several hundred examples with a usable number of failures, which is marginal but workable. Framing it per concept pair pools evidence across all participants and is limited by the number of concept pairs rather than the number of children.

The decision was to move the unit of analysis from the student to the concept pair for the primary research question, and to the attempt for the secondary one. This is the single most consequential methodological decision in the thesis, and it is what makes the result reported in Section 6.1.4 defensible with 19 participants.

4.3.3 Model Family: Graph Attention Versus Classical Alternatives

Graph attention was selected over graph convolution and neighbour sampling for a specific reason rather than a general preference. The deployed heuristic already contains a hand-set attention weight: personal confusion evidence is weighted three times pooled cohort evidence. That constant encodes a judgement that a child's own evidence should outweigh the cohort's, with the ratio guessed rather than learned. Attention learns that ratio per concept pair from data, which is a concrete and testable motivation.

Four classical alternatives were nonetheless implemented as baselines rather than dismissed: Bayesian Knowledge Tracing, which is purpose-built for per-concept mastery estimation and works at very low data volumes; a Rasch one-parameter item response model, which is the standard approach in adaptive testing; a concept difficulty prior; and a popularity prior. The decision to implement all of them before building the graph model proved correct, and Section 6.2 reports why.

4.3.4 Similarity Features: Computed Versus Learned

Children confuse apple with cherry because the images resemble one another, and cat with cap because the words sound alike. The concept nodes as originally designed encoded neither property, carrying only a label, a category, and a sequence index. Two options were available: allow the graph model to learn concept representations

from interaction data alone, or compute similarity features explicitly and attach them to the graph.

The decision was to compute them explicitly, for two reasons. With 42 concepts touched out of 96, learned embeddings would be undefined for most of the curriculum, whereas computed features are available for every concept including those no child has attempted. And an explicit feature is testable: it is possible to ask whether visual similarity predicts confusion, which is a scientific question, in a way that it is not possible to ask the same of an uninterpretable learned embedding. Section 6.1.4 reports the answer, and it is the strongest positive finding in this thesis.

4.4 Detailed Design

4.4.1 Tier Progression and Mastery

Each concept is modelled as a state machine with three tier states and a terminal learnt state. Transition from Tier 1 to Tier 2 requires that the Tier 1 mastery criterion be met. A concept reaches the learnt state when both Tier 1 and Tier 2 are passed; Tier 3 is enrichment and does not gate the learnt state, because its colouring activity is deliberately unscored.

Mastery within a tier is evaluated per block. A block comprises three scored trials, and the criterion requires two correct. The score recorded against the tier is the proportion correct within the assessed block, taking values of 0, one third, two thirds, or 1, which is why the score distributions reported in Section 6.1.1 are discrete at those four points.

4.4.2 Adaptive Difficulty and the Relearn Pathway

Difficulty within a tier is expressed as the number of distractors presented alongside the target. The controller increases the distractor count after consecutive correct responses and reduces it after consecutive incorrect responses. The adjustment is deliberately asymmetric, falling faster than it rises, because the cost of a difficulty spike for an autistic learner is a frustration response rather than a lower score.

Within-block error handling is graduated and non-punitive. The first incorrect attempt produces a subtle visual indication. The second removes a distractor. The

third highlights the correct answer with a hint. There is no failure state presented to the child at any point.

Where a block fails to meet the mastery criterion, the relearn pathway activates across three levels that vary the stimulus rather than repeating it. Level 1 switches the visual stimulus from real photograph to illustrated cartoon. Level 2 introduces a short contextual animation before the retry. Level 3 returns the child to the full familiarisation screen with animation-on-tap. The design principle is that a child who failed to recognise a photograph is unlikely to be helped by being shown the same photograph again, and that varying the modality tests whether the difficulty lies in the concept or in its presentation. Exhausting Level 3 without meeting the criterion raises the teacher handover flag rather than restarting the cycle.

4.4.3 Graph Knowledge Base Schema

The graph comprises four node labels and thirteen relationship types. Student, Concept, and Category nodes form the structural layer, and Attempt nodes form the temporal layer added during this work.

Table 4.2: Graph Knowledge Base relationship types

Relationship	From	To	Semantics
HAS_CONCEPT	Category	Concept	Static curriculum structure
T1_SCORE	Student	Concept	Tier 1 mastery score, attempts, pass state
T2_SCORE	Student	Concept	Tier 2 mastery score, attempts, pass state
ACTIVITY_SCORE	Student	Concept	Category game score and difficulty level
T1_ENGAGEMENT	Student	Concept	Tap count, dwell time, image format used
T2_ENGAGEMENT	Student	Concept	Tap count, dwell time
CONFUSION	Concept	Concept	Tier 1 identification confusion, pooled across students
T2_NAME_CONFUSION	Concept	Concept	Tier 2 name confusion,

Relationship	From	To	Semantics
			pooled across students
ACTIVITY_CONFUSION	Concept	Concept	Activity confusion, scoped to one student
ACTIVITY_FORMAT_CONFUSION	Concept	Concept	Format-specific confusion, scoped to one student
VISUAL_SIMILARITY	Concept	Concept	Computed visual resemblance score
PHONETIC_SIMILARITY	Concept	Concept	Computed phonetic resemblance score
ATTEMPTED / ON	Student / Attempt	Attempt / Concept	Append-only attempt history

The distinction between pooled and student-scoped confusion edges is significant. Pooled edges accumulate evidence across the cohort and support cold-start reasoning for a child who has not yet encountered a concept. Student-scoped edges carry the individual's own evidence. The deployed heuristic blends the two, weighting personal evidence three times the pooled value, which is the hand-set constant discussed in Section 4.3.

4.4.4 Distractor Selection Chain

Distractor selection operates as an ordered fallback chain, and the chain structure is what allows new intelligence to be introduced without risk. The first source queries the Graph Knowledge Base for confusion edges under a 400 millisecond timeout. The second falls back to failure records in the PostgreSQL log. The third falls back to the adjacent concepts in the curriculum sequence. Each stage degrades to the next on timeout, error, or empty result.

The third stage is implicated in the methodological finding reported in Section 6.2.1 and is therefore worth stating precisely: when no confusion evidence exists for a concept, which is true for most concepts, the system presents the next concepts in the curriculum sequence as distractors. This is a reasonable engineering default and a serious evaluation hazard, for reasons set out in that section.

4.4.5 Concept Similarity Feature Layer

Visual similarity is computed offline by passing each concept's real-photograph asset through a pretrained convolutional image encoder and taking the pairwise cosine similarity between the resulting embeddings. Phonetic similarity is computed from the concept labels using phoneme-level overlap and edit distance. Both are computed for within-category pairs only, on the grounds that the system never presents a cross-category distractor, and both are written to the graph as weighted relationships.

The snapshot analysed in Chapter 6 contains 462 visual similarity relationships over 286 undirected pairs, and 534 phonetic similarity relationships over 320 undirected pairs. Critically, these features are defined for concepts that no child has ever attempted, which is what makes them available at cold start.

4.5 Standards and Professional Practice

The instructional design follows Applied Behaviour Analysis principles as described in Section 2.2, specifically task decomposition into sequentially mastered steps, reinforcement to consolidate learning, and an explicit mastery criterion before progression.

Interface design follows established guidance for autism-specific user experience, specifically the properties of being structured, predictable, customisable, sense aware, attention retaining, and frustration free. Learning screens lock to landscape, layout reserves space so that interactive content does not shift under the child's finger, and break prompts never appear mid-interaction.

Engineering practice follows the conventions of the wider project. The API is documented with an OpenAPI specification. Routes are guarded by token verification, a role guard, and an ownership check confirming that the requesting teacher is assigned to the student whose data is being accessed. Rate limits are applied separately to authentication, concept, and live session routes. Schema changes are applied through versioned migrations, and automatic schema synchronisation is disabled in production.

Two practices specific to responsible AI are enforced in code rather than policy. The language model prompt layer is maintained as a separate configuration with explicit instructions for plain-language, teacher-addressed, non-diagnostic output. The graph inference path carries a kill switch that returns the system to purely heuristic behaviour instantly.

5 IMPLEMENTATION, INTEGRATION AND TESTING

5.1 Development Environment and Technology Stack

The mobile application is implemented in React Native 0.81 with React 19 on Expo SDK 54, targeting Android tablets as the primary platform. Navigation uses React Navigation 7, state is managed with a combination of a lightweight global store for authentication and React context for the learning session, and networking uses Axios. Concept activity rendering uses react-native-svg for interactive graphics and the colouring canvas, Reanimated and Lottie for animation, and expo-audio for narration playback. Orientation locking uses expo-screen-orientation, and sensory preferences persist through the student record rather than device storage so that they follow the child across devices.

The backend API is implemented in Node.js with Express 5 and Sequelize 6 over PostgreSQL, hosted on managed cloud infrastructure. Authentication uses signed tokens with password hashing. Security middleware comprises standard header hardening, cross-origin restriction, request validation, and route-specific rate limiting. API documentation is generated from source annotations.

The Graph Knowledge Base service is implemented in Python with FastAPI over the official Neo4j driver, connecting to a managed Neo4j instance. The connection pool is capped and connections are recycled on a fixed interval to suit the constraints of the database tier in use. The research code for the graph model, the dataset builder, the baselines, and the evaluation harness is maintained in the same repository but is never imported by the serving application, and its dependencies are specified separately so that the production container does not carry the machine learning stack.

Teacher-facing interpretive summaries are generated through a hosted large language model API, with prompts assembled from anonymised session statistics and graph-derived analytics.

5.2 Component Implementation

5.2.1 Tier Engine

The tier engine is implemented in the concept service as a set of pure evaluation functions over the current block state, separated from the persistence layer. This separation was made for testability: the mastery decision, the difficulty adjustment, and the relearn level selection are each a function of block state and configuration constants, and each is unit tested against enumerated scenarios without a database.

The mastery constant is defined in one place and referenced throughout, so that the threshold cannot drift between tiers. Difficulty adjustment and relearn progression are likewise driven by named constants rather than inline values, which is what made the configuration auditable during the analysis reported in Chapter 6.

5.2.2 Interaction Logging

Every interaction writes a structured event to the interaction log table with the student, category, concept, tier, event type, and a flexible payload. The payload carries the response, the selected option where the response was incorrect, the active difficulty, the elapsed time, and, for activity events, the full set of options presented.

That last field proved decisive. During the Phase 1 evaluation it emerged that only the activity attempt event recorded the options presented to the child, while the four other scored attempt types did not. The consequence is analysed in Section 6.2.1. The remediation, extending option set logging to all scored attempt types, requires no schema change because the payload column is schemaless, and is recorded in Section 9.4 as the first required step before the ranking evaluation can be repeated validly.

5.2.3 Graph Synchronization

Graph writes are issued from the concept service after each scored block, in a fire-and-forget pattern that does not await the response. A failure to write to the graph is logged and otherwise ignored, because the authoritative record is PostgreSQL and the graph is a projection of it. This means the graph can be rebuilt from the event log at any time, which is precisely what the dataset builder does.

Confusion edge weights are incremented rather than overwritten, so pooled confusion accumulates across the cohort. Score and engagement edges were originally written with overwrite semantics, which is correct for a serving-time feature store but destroys history. The addition of append-only attempt nodes, described in Section 4.3, resolves this without changing the serving semantics: the current-state edges remain fast to query, and the attempt nodes retain the history the edges discard.

5.2.4 Similarity Feature Pipeline

The similarity pipeline runs offline and writes into the graph. The visual stage loads each concept's photograph asset, computes an embedding with a pretrained image encoder, and writes pairwise cosine similarity for within-category pairs. The phonetic stage derives a phoneme sequence from each concept label and computes a normalised overlap score. Both stages are idempotent and are re-run when concepts are added. The pipeline is not invoked at request time and has no production dependency.

5.2.5 Teacher-Facing Analytics

The teacher interface presents concept progress at three levels. The category level shows progress across the nine curriculum categories. The concept level shows tier completion status, mastery scores with the difficulty level active at the time of scoring, relearn cycle counts, and audio replay frequency. The inline summary presents a short plain-language interpretation generated from the session data, with graph-derived context where available.

Summaries are generated on demand rather than on a schedule, and generation is isolated from the child-facing session path. If generation fails or is slow, the numeric dashboard renders without it and the interface states that the summary is unavailable rather than displaying a placeholder or a fabricated value.

5.3 Integration with the other components

Integration with the handwriting, pronunciation, and dialogue components occurs at four surfaces.

The authentication and authorisation layer is shared. All concept routes are guarded by the same token verification, role guard, and student ownership check used by the other modules, so that a teacher can only read or write data for children assigned to them.

The student and teacher data model is shared. The Concept Learning Module contributes its own progress, interaction log, and report tables but references the common student and teacher records, which means a child added by the principal is immediately available across all four modules without per-module registration.

The Teacher Workspace is shared. The student profile presents module progress as four tabs, and the concept tab conforms to the same card layout, summary structure, and report entry points as the other three. This was coordinated across the group so that a teacher learns one reporting interface rather than four.

The sensory preference layer is shared. The avatar, theme, and font preferences captured in the one-time initial assessment are stored once against the student record and applied by all four modules, which is what makes the interface consistent across a session that moves between modules.

Deployment is also shared: all services are containerised and run as a single composed stack behind one reverse proxy, with only the proxy exposed publicly and the backend and Python services reachable only on the internal network.

5.4 Verification and Software Testing

Verification was conducted at three levels.

Unit testing covers the pure evaluation functions of the tier engine, the validation layer, and the utility and formatting functions, on both the backend and the mobile application. The functions selected for unit testing are those where a defect would produce a silently wrong instructional decision rather than a visible error, specifically mastery evaluation, difficulty adjustment, and relearn level selection.

Integration testing covers the API surface, exercising the concept routes against a test database with the authorisation guards active, including negative cases confirming that a teacher cannot access a student not assigned to them.

Scenario testing covers the adaptive behaviours end to end, with expected outcomes defined in advance and compared against actual system behaviour. Table 5.1 summarises the scenarios verified.

Table 5.1: Adaptive behaviour verification scenarios

Scenario	Expected behaviour	Result
Two correct of three in a block	Tier marked passed; progression permitted	Pass
One correct of three in a block	Tier not passed; relearn Level 1 triggered	Pass
Relearn Level 1 failure	Stimulus switches to animation at Level 2	Pass
Relearn Level 3 exhausted	Teacher handover flag raised; no further loop	Pass
Consecutive correct answers	Distractor count increases	Pass
Consecutive incorrect answers	Distractor count reduces	Pass
First incorrect attempt in block	Subtle visual indication only	Pass
Third incorrect attempt in block	Correct answer highlighted with hint	Pass
Tier 2 attempted without Tier 1 pass	Progression blocked by tier gate	Pass
Graph service unavailable	Distractor chain falls back to log, then sequence	Pass
Graph service exceeds 400 ms	Timeout; fallback applied; no child-visible delay	Pass
Language model generation fails	Dashboard renders without summary; no fabricated value	Pass
Repeated audio replay request	Replay served; replay count logged	Pass
Session resumed after interruption	Child resumes at last reached tier	Pass

The fallback scenarios in the lower half of that table were given particular attention, because the correctness property that matters most in this system is not that the

intelligence layer produces a good answer but that its absence is invisible to the child.

5.5 Deployment and Operation

The production system runs as a composed container stack on a single virtual machine. A reverse proxy terminates TLS and is the only publicly exposed service. The backend API and the three Python services, including the Graph Knowledge Base service, are reachable only on the internal container network. Container images are built and published automatically from the main branch, and a single service can be rolled back independently by pinning its image tag.

External managed services comprise the PostgreSQL database, blob storage for audio and image assets, the Neo4j graph database, and the language model API. The application timezone is fixed to Asia/Colombo, which matters because daily practice limits and session dates are computed in local time.

The graph inference path is governed by an environment flag that disables it entirely, returning the system to heuristic behaviour. This was specified as a safety requirement in Section 3.6 and is the mechanism by which the research model could be deployed in shadow mode without risk to the child-facing experience.

Operationally, the module was running in the participating school throughout the evaluation period, with sessions conducted by teachers during normal timetabled activity rather than in researcher-supervised sessions. The data reported in Chapter 6 is therefore production data from ordinary use, not data from a controlled laboratory condition.

6 RESULTS AND VALIDATION

This chapter reports the evaluation of the Concept Learning Module following deployment at the participating special education school. All figures derive from a snapshot of the production knowledge graph taken on 11 October 2026, comprising 570 nodes and 2,663 relationships, together with the versioned interaction datasets described in Section 3.7. Two non-participant accounts, one system test account and one account carrying no participant identity, are excluded from every analysis reported here, leaving 19 participant records with tier-level learning data from the 21 child records in the graph.

6.1 Research Results

6.1.1 Deployment Scale and Coverage

The module recorded 441 scored attempts from 18 participants across 42 distinct concepts over 27 active days between 2 May 2026 and 9 October 2026. Of these attempts, 200 were Tier 1 identification trials, 169 were Tier 2 word recognition trials, and 72 were category activity trials. The overall attempt-level pass rate was 76.4 per cent, distributed as 78.0 per cent at Tier 1, 75.2 per cent at Tier 2, and 75.0 per cent at activity level.

Attempt volume remained concentrated in a subset of the curriculum. The fruits category accounted for 283 of 441 attempts, or 64.2 per cent, with animals contributing 93, household 24, classroom 14, colors 12, shapes 8, family 5, and numbers 2. Three of the nine categories recorded no interaction at all. This distribution is a material limitation on generalisation and is examined in Section 7.5.

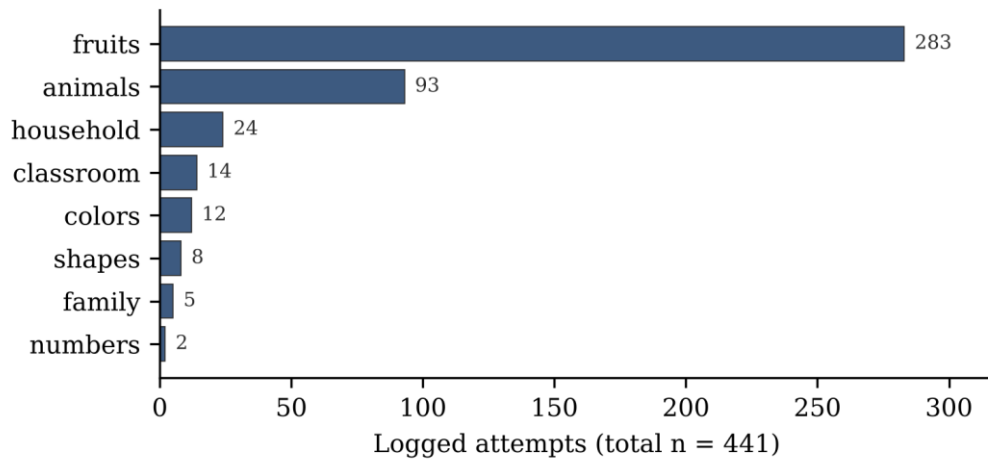


Figure 6.1: Distribution of logged attempts across curriculum categories

Engagement with the module was uneven across participants, which is expected given that session frequency was determined by the teacher and the school timetable rather than by a research protocol. Attempt counts per participant ranged from 2 to 55.

6.1.2 Tier Mastery Outcomes

The graph holds 127 Tier 1 mastery records and 120 Tier 2 mastery records across 19 participants and 41 concepts. Tier 1 was passed in 102 of 127 records, a pass rate of 80.3 per cent, with a mean score of 0.790 and a standard deviation of 0.308. Tier 2 was passed in 118 of 120 records, a pass rate of 98.3 per cent, with a mean score of 0.881 and a standard deviation of 0.172.

Scores are discrete at four values because mastery is assessed over a three-trial block, as described in Section 4.4.1. The distribution of those values is shown in Figure 6.2.

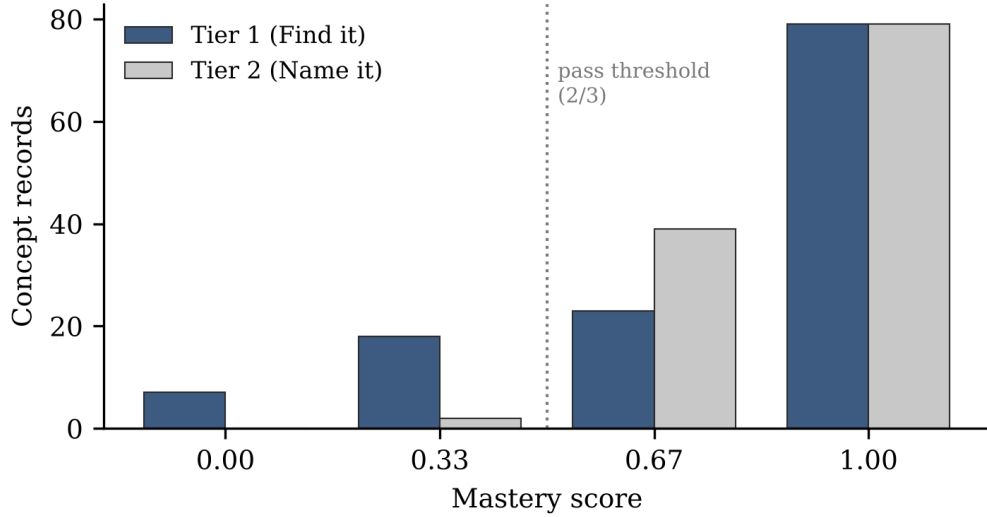


Figure 6.2: Distribution of Tier 1 and Tier 2 mastery scores against the pass threshold

Two features of that distribution are informative. Tier 1 produced seven records at a score of zero and eighteen at one third, whereas Tier 2 produced no records at zero and only two at one third. The lower tail therefore disappears almost entirely at Tier 2. The reason is the tier gate rather than an intrinsic property of the task: of the 102 concepts passed at Tier 1, 95 went on to a Tier 2 attempt, and 98.95 per cent of those passed. Children reached Tier 2 having already demonstrated recognition of the concept, which is the behaviour the gated architecture was designed to produce.

A paired comparison restricted to the 118 student-concept pairs present at both tiers gives a mean Tier 1 score of 0.791 and a mean Tier 2 score of 0.881, a difference of 0.090 with a 95 per cent confidence interval of 0.035 to 0.146 and a paired t statistic of 3.198 on 117 degrees of freedom, $p = 0.0018$. The interpretation of this difference requires care and is addressed in Section 7.1, because it is a consequence of tier gating and cannot be read as evidence that naming is easier than identification.

Twenty-five Tier 2 records exist without a corresponding Tier 1 pass. These are historical records predating the current gate enforcement, and they are retained in the dataset rather than removed because silently deleting inconvenient records would misrepresent the deployment history.

Per-participant outcomes are summarised in Table 6.1, ordered by Tier 1 record count. Participants are identified by anonymised index rather than by name.

Table 6.1: Per-participant mastery outcomes

P	T1 records	T1 passed	T1 mean	T2 records	T2 passed	T2 mean	Concepts learnt
P01	18	18	0.944	17	17	0.941	17
P02	18	12	0.741	19	19	0.895	11
P03	17	16	0.882	17	17	0.902	16
P04	13	8	0.692	12	12	0.917	8
P05	11	10	0.818	11	10	0.758	9
P06	9	7	0.630	5	5	0.800	3
P07	8	6	0.792	7	6	0.857	6
P08	7	4	0.571	7	7	0.905	4
P09	5	5	1.000	5	5	0.933	5
P10	4	4	1.000	4	4	1.000	4
P11	3	3	1.000	2	2	1.000	2
P12	3	2	0.667	3	3	0.667	2
P13	2	1	0.500	2	2	0.833	1
P14	2	2	0.833	2	2	1.000	2
P15	2	1	0.667	2	2	0.667	1
P16	2	0	0.167	2	2	0.667	0
P17	1	1	1.000	1	1	1.000	1
P18	1	1	1.000	1	1	1.000	1
P19	1	1	0.667	1	1	0.667	1

Across the 19 participants, a mean of 4.95 concepts and a median of 3 concepts reached the learnt state, requiring passes at both Tier 1 and Tier 2. The range extends from 0 to 17. The spread reflects differences in exposure as much as differences in ability: the three participants with the highest concept counts are also the three with the most recorded sessions.

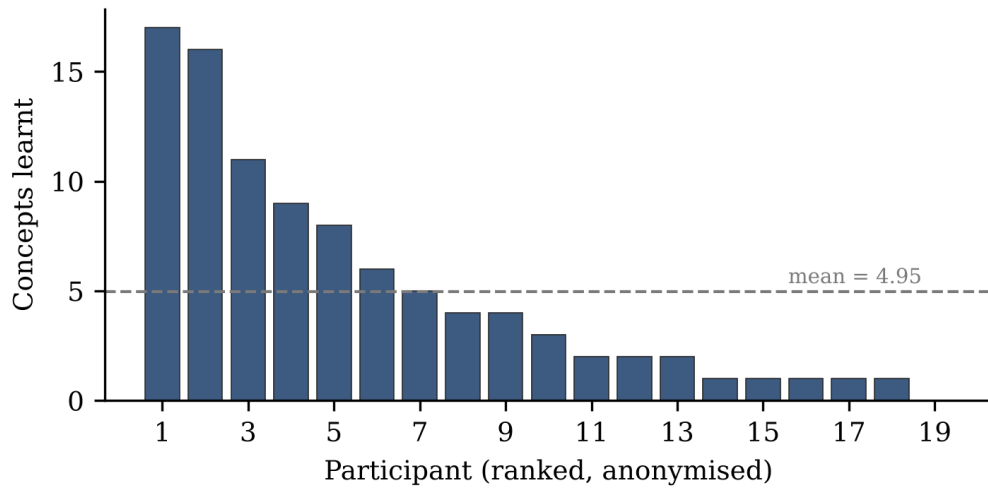


Figure 6.3: Concepts reaching the learnt state per participant

6.1.3 Within-Concept Learning Gain

The primary learning outcome measure is the within-concept gain defined in Section 3.4, computed as the difference between the first and last recorded score for a given participant, concept, and tier where more than one attempt exists. This is a within-subject measure, and it is the measure on which the learning claim of this thesis rests.

Of 299 participant-concept-tier sequences, 69 contain more than one scored attempt, with a maximum of 12 attempts in a single sequence and a mean of 1.475 attempts per sequence. Across those 69 repeated sequences the mean gain was 0.358 score points with a standard deviation of 0.509, giving a 95 per cent confidence interval of 0.235 to 0.480. Forty-seven sequences improved, 14 remained unchanged, and 8 declined.

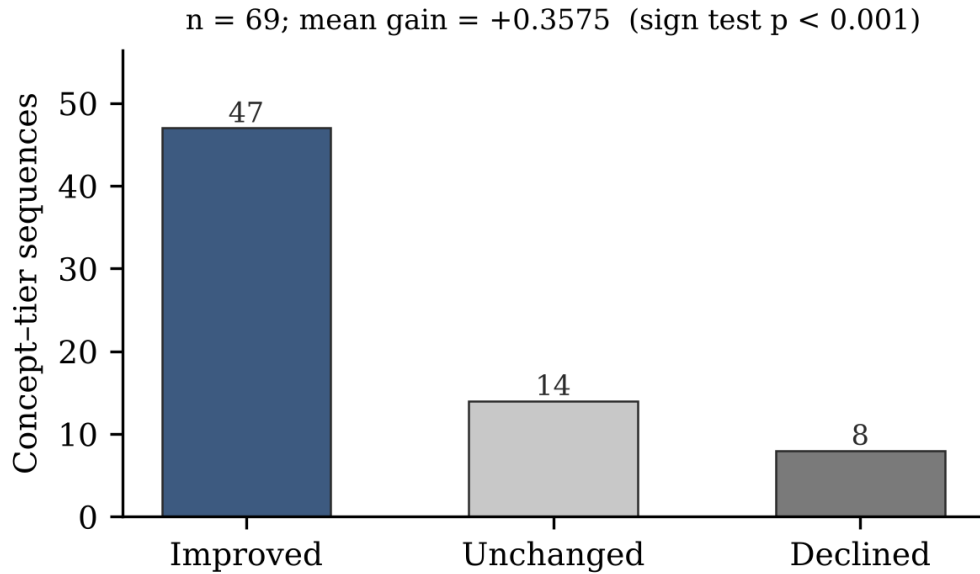


Figure 6.4: Direction of within-concept change across repeated exposures

A two-sided sign test on the 55 sequences that changed in either direction gives $p = 8.07 \times 10^{-8}$. The improvement is therefore reliable and not attributable to chance. It should be read as evidence that repeated structured exposure within the module is associated with improved performance on the same concept, which is a within-system learning effect. It is not evidence of transfer to contexts outside the module, which this design cannot assess.

6.1.4 Concept Similarity and Observed Confusion

This section reports the principal research finding of the thesis. The question, stated as RQ2 in Section 1.6, is whether computed concept similarity predicts the confusions children actually make.

The knowledge graph records 44 undirected concept pairs confused during Tier 1 identification and 52 during Tier 2 name recognition, giving 70 distinct pairs observed as confused at least once. The similarity feature layer scores 286 undirected within-category pairs for visual similarity and 320 for phonetic similarity. The analysis compares similarity scores between pairs that were observed as confused and pairs that were not.

For visual similarity, the 48 confused pairs carried a mean score of 0.858 and a median of 0.895, against a mean of 0.804 and a median of 0.777 for the 238 pairs not observed as confused. A Mann-Whitney U test gives $U = 7713$, $z = 3.828$, $p = 0.00013$, with a rank-biserial correlation of 0.350 and Cohen's d of 0.623. Visual similarity is therefore significantly higher among confused pairs, with a medium effect size.

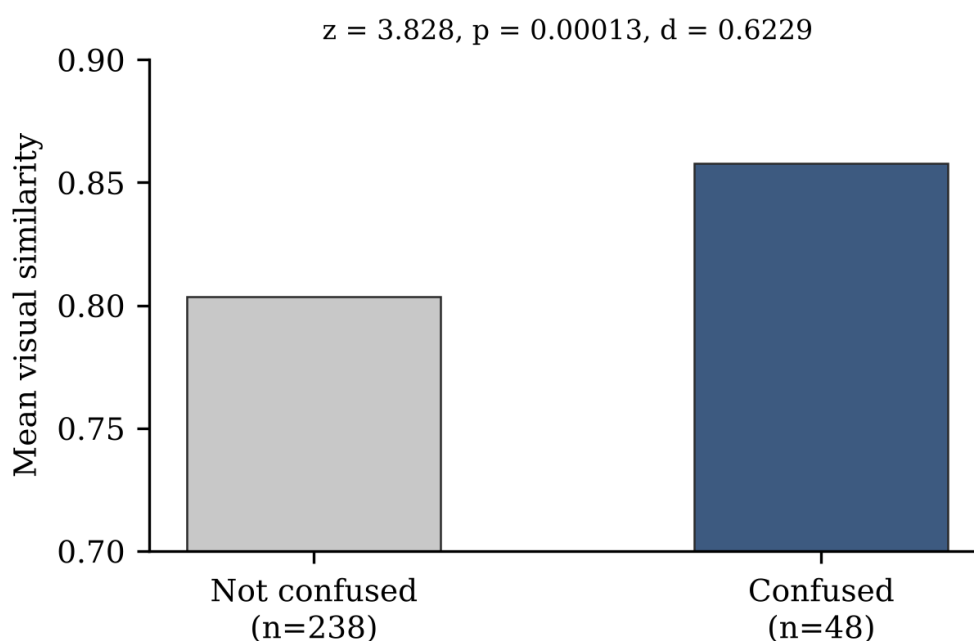


Figure 6.5: Visual similarity for concept pairs observed as confused against pairs not confused

For phonetic similarity, the 49 confused pairs carried a mean score of 0.119 and a median of 0.108, against a mean of 0.107 and a median of 0.089 for the 271 pairs not confused. The same test gives $z = 1.346$, $p = 0.178$, with a rank-biserial correlation of 0.121 and Cohen's d of 0.139. The difference is in the expected direction but is not statistically reliable, and the effect size is negligible.

Table 6.2: Similarity scores for confused against non-confused concept pairs

Feature	Confused (n, mean)	Not confused (n, mean)	z	p	Cohen's d
Visual similarity	48, 0.858	238, 0.804	3.828	0.00013	0.623
Phonetic similarity	49, 0.119	271, 0.107	1.346	0.178	0.139

The practical consequence is specific. The distractor selection chain described in Section 4.4.4 should prefer visual similarity over phonetic similarity as its evidence source when no observed confusion edge exists for a pair, and the phonetic feature should not be weighted in that decision on the strength of the present evidence.

The most consequential property of this finding concerns cold start. Of the 96 concepts in the curriculum, 42 have been attempted by at least one child and 54 have not. Of those 54 untouched concepts, 52 carry similarity relationships, a coverage of 96.3 per cent. The similarity layer therefore supplies a principled, empirically validated basis for distractor selection on almost the entire curriculum that no child has yet encountered, which is precisely where the per-item approach used by existing systems has nothing to offer and falls back to an arbitrary sequence position.

The most strongly weighted observed confusions are consistent with the visual account. At Tier 1 the highest weighted pairs are apple with cherry at weight 21, apple with banana at 20, and grapes with mango at 15. At Tier 2 name recognition the highest weighted pairs are apple with banana, apple with cherry, grapes with mango, and orange with papaya, each at weight 7, followed by red with green at 6. The appearance of red with green among the name confusions, where the written words share no visual resemblance, indicates that Tier 2 confusion is not purely a function of word form and that a separate account is required for name-level errors.

6.1.5 Engagement and Scaffolding Reliance

Engagement measures were recorded for 123 Tier 1 concept records and 113 Tier 2 records. At Tier 1 the mean tap and audio replay count was 4.97 with a median of 2, and the median dwell time was 2.37 seconds with a mean of 4.86 seconds. At Tier 2 the mean tap count fell to 1.59 with a median of 2, and the median dwell time fell to 1.35 seconds with a mean of 3.95 seconds.

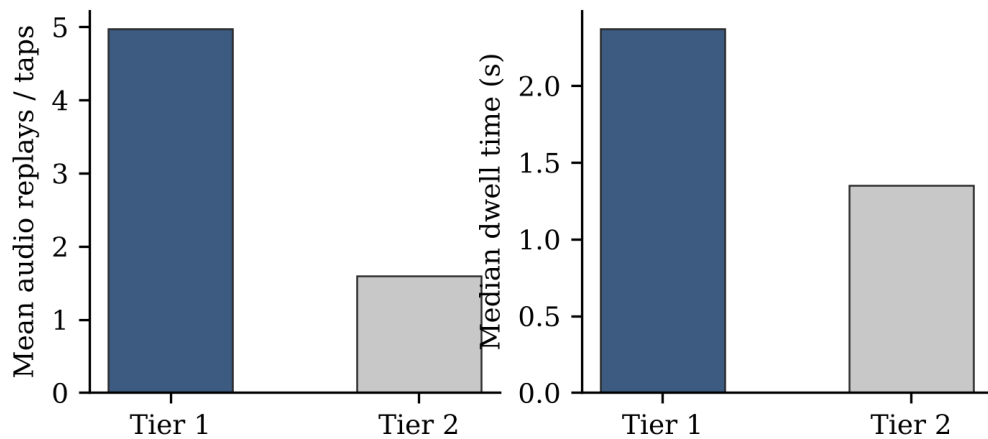


Figure 6.6: Audio replay and dwell time at Tier 1 compared with Tier 2

The reduction in replay frequency between tiers is consistent with the design intention that auditory scaffolding is most needed during initial familiarisation and is drawn on less once the concept is recognised. The mean values at both tiers substantially exceed the medians, indicating a small number of records with very high replay counts, which is the pattern the teacher dashboard surfaces as persistent auditory support reliance. All Tier 1 engagement records used the real photograph format, confirming that the configured default stimulus was applied as specified and that the relearn pathway's alternative formats were reached only on failure rather than by default.

6.1.6 Category Activity Outcomes

Category activities, which unlock as concepts within a category reach the learnt state, recorded 48 scored results from 7 participants, with a pass rate of 79.2 per cent and a mean score of 0.708. The difficulty ladder was exercised across levels 1 to 4, with 23 results at level 1, 8 at level 2, 4 at level 3, and 1 at level 4, and 12 results carrying no recorded difficulty level. The concentration at level 1 reflects that most participants reached the activity stage late in the deployment period and had limited opportunity to ascend the ladder. The 12 results without a difficulty level are a data quality defect rather than a finding, and are recorded in Section 9.4 as requiring correction.

6.2 Product and Technical Results

This section reports the evaluation of the graph intelligence strand, which addresses RQ3. The result is negative, and the reasons are methodologically significant.

6.2.1 The Exposure Policy Confound

The distractor ranking task was evaluated under leave-one-student-out cross-validation against the full baseline suite. The outcome is reported in Table 6.3, and it initially appears to show that a trivial sequential scorer substantially outperforms every alternative including the production heuristic.

Table 6.3: Distractor ranking under leave-one-student-out cross-validation (confounded)

Scorer	Hits@1	Hits@2	Hits@2 95% CI	MRR
sequential	0.405	0.888	[0.844, 0.932]	0.663
heuristic (production)	0.298	0.639	[0.576, 0.707]	0.560
personal only	0.351	0.580	[0.512, 0.649]	0.541
cohort only	0.239	0.551	[0.483, 0.620]	0.498
popularity	0.239	0.434	[0.366, 0.498]	0.451
GCN	0.172	0.377	[0.313, 0.443]	0.399
GraphSAGE	0.147	0.307	[0.245, 0.368]	0.360
GATv2	0.119	0.236	[0.181, 0.294]	0.327
random	0.117	0.171	[0.122, 0.224]	0.294

That table must not be read as evidence that a sequential scorer is the best model. It is measuring the data collection policy rather than confusability.

The mechanism is as follows. The third stage of the distractor selection chain, described in Section 4.4.4, returns the next concepts in the curriculum sequence whenever the knowledge graph holds no confusion edge for a pair, which is true for most pairs because most concepts have no interaction history. A child can only select a distractor that was presented. In the earlier dataset version on which this analysis was first run, 97.0 per cent of observed confusions were with a concept exactly one or two positions ahead of the target in the category sequence. The sequential scorer

therefore scores highly because it is predicting the system's own fallback behaviour, and the production heuristic scores worse precisely because it sometimes ranks a genuinely confusable concept above the two the child was actually offered, which the child had no opportunity to select.

The label against which distractor quality is being measured is constructed by the policy under evaluation. Compounding this, only 18 of 732 attempts in that dataset recorded the option set presented, because the option set field was populated by the activity attempt event alone and not by the four other scored attempt types. For 97.5 per cent of the data the candidate set cannot be reconstructed, and therefore no unbiased ranking metric can be computed, for the heuristic or for any model.

A deeper form of the same problem applies to the confusion dataset generally. Concept pairs never presented together can never be observed as confused, so the negative examples are not true negatives but unobserved ones. This is positive-unlabelled learning under a biased exposure policy.

The conclusion is that no offline claim about distractor quality is falsifiable with the present instrumentation. Reporting the GAT as the worst performer in Table 6.3 would be as unjustified as reporting the sequential scorer as the best. This finding is the reason the graph model was never promoted out of shadow mode, and identifying it cost two analysis scripts rather than the three weeks of model development that would otherwise have produced an unfalsifiable headline metric.

6.2.2 Attempt Outcome Prediction

The attempt outcome task is not subject to the same confound, because the outcome of an attempt is observed directly rather than inferred from a constructed candidate set. Results are reported in Table 6.4.

Table 6.4: Attempt outcome prediction under leave-one-student-out cross-validation

Model	Accuracy	Acc 95% CI	Log loss	Brier	ROC-AUC
Rasch (1PL)	0.750	[0.693, 0.804]	0.612	0.199	0.504
majority	0.747	[0.685, 0.805]	0.572	0.191	0.368
concept	0.745	[0.685, 0.803]	0.583	0.195	0.505

Model	Accuracy	Acc 95% CI	Log loss	Brier	ROC-AUC
difficulty					
BKT	0.730	[0.662, 0.795]	0.568	0.191	0.558
GraphSAGE	0.653	[0.543, 0.746]	0.662	0.233	0.511
GCN	0.642	[0.563, 0.715]	0.654	0.230	0.521
GATv2	0.607	[0.518, 0.688]	0.678	0.242	0.500

The base rate is 0.747, so every model sits at or below the majority floor on accuracy, and this is why accuracy was excluded as a primary metric in Section 3.4. The discriminative metric tells the real story: ROC-AUC values cluster around 0.50 for every model except Bayesian Knowledge Tracing at 0.558, which means there is essentially no discriminative signal. The Rasch model, which is the standard approach for exactly this task, separates the classes no better than chance at 0.504. The three graph architectures perform worst of all, with the GATv2 configuration at exactly 0.500, and they also carry the highest log loss and Brier scores, indicating that they are not merely uninformative but poorly calibrated.

This is not a defect in the implementation. It indicates that with the available volume of attempts across 19 children and 42 touched concepts, per-attempt correctness is not predictable from student ability and item difficulty alone. The honest conclusion is that the data does not yet support this task, and that the graph architectures in particular are overfitting a dataset far below the scale at which such models are normally applied.

6.2.3 Mastery Decision Prediction

A third task asked whether Tier 1 performance predicts Tier 2 outcome. Across 103 outcomes over 103 groups, Bayesian Knowledge Tracing, a fixed two-thirds threshold rule, and a majority predictor all returned an identical accuracy of 0.728 with confidence intervals of [0.651, 0.790], and ROC-AUC values of 0.534, 0.476 and 0.359 respectively. No approach separates the classes. Given the tier gating effect described in Section 6.1.2, under which almost every child reaching Tier 2 has

already passed Tier 1, this is the expected outcome: the gate removes most of the variance that such a predictor would need.

6.2.4 Summary of the Graph Intelligence Result

The Graph Attention Network did not outperform the deployed heuristic or any classical baseline on any of the three tasks evaluated. On the ranking task the comparison is not valid at all, for the reasons given in Section 6.2.1. On the two prediction tasks the comparison is valid, and the graph models lose to simple baselines that themselves perform at close to chance.

RQ3 is therefore answered in the negative. The phase gate defined for this strand of the work specified that the model should not influence what a child sees unless it demonstrably beat the heuristic. It did not, and it was not promoted. The deployed system continues to use the rule-based engine, and the graph layer contributes through its feature and serving roles rather than through learned inference.

6.3 User and Stakeholder Validation

Qualitative validation was conducted with the special education teachers who operated the system. Teachers interacted with both the Teacher Workspace and the Student Workspace and provided feedback through task-based assessment and guided discussion, focusing on whether the interface was clear to navigate, whether the generated summaries and reports were understandable and useful in practice, whether the teacher handover mechanism supported their professional workflow, and whether the overall design was appropriate for use in a special education classroom.

Child-facing evaluation was observational and conducted under the direct supervision of the assigned teacher, with data collection limited to engagement, task completion, and any visible signs of distress or disengagement as identified by that teacher. No session was continued where the supervising teacher judged that it should stop.

Three themes emerged consistently from teacher feedback and each produced a design change. Teachers valued the teacher handover flag specifically because it bounded the number of consecutive failures a child could experience, confirming the

design rationale given in Section 2.2. Teachers required that generated summaries describe rather than interpret, which reinforced the non-diagnostic prompt constraint recorded in Section 3.6. Teachers asked for concepts outside the original nine categories, which is why a family category was created during the deployment period and appears in the graph without a category relationship.

A quantified usability instrument was not administered within the evaluation period. The qualitative findings reported here are therefore indicative rather than measured, and the absence of a System Usability Scale score is recorded as a limitation in Section 7.5 and as required work in Section 9.4.

6.4 Integration Results

Integration with the three other components was verified across the four shared surfaces described in Section 5.3. Authentication, role guarding, and student ownership checks operate identically across all concept routes, and negative test cases confirmed that a teacher cannot retrieve concept data for a child not assigned to them. Students created once by the principal are available across all four modules without per-module registration. The concept progress tab conforms to the shared Teacher Workspace layout. Sensory preferences captured in the one-time initial assessment are applied by all four modules from a single stored record.

The fallback behaviour that matters most for integration was verified explicitly: with the graph service stopped, concept sessions continued without child-visible change, falling through to the log-based and then sequence-based distractor sources. With the graph service artificially delayed beyond the 400 millisecond timeout, the same fallback applied and no delay was observable in the child-facing interface. With language model generation failing, the teacher dashboard rendered its numeric content and stated that the summary was unavailable rather than displaying a placeholder value.

6.5 Summary of Findings

The findings of this chapter are the following.

- The module was deployed and used in live classroom practice with 21 registered children, 19 of whom produced tier-level learning data, generating 441 scored attempts across 42 concepts over 27 active days.
- Tier 1 mastery was achieved in 80.3 per cent of concept records and Tier 2 in 98.3 per cent. The tier gate functions as designed: of concepts passed at Tier 1 and subsequently attempted at Tier 2, 98.95 per cent passed.
- Within-concept performance improved reliably across repeated exposures, by a mean of 0.358 score points with a 95 per cent confidence interval of 0.235 to 0.480, with 47 sequences improving against 8 declining (sign test $p < 0.001$).
- Visual similarity between concepts significantly predicts the confusions children actually make ($z = 3.83$, $p = 0.00013$, $d = 0.62$). Phonetic similarity does not ($p = 0.178$, $d = 0.14$).
- The similarity layer covers 96.3 per cent of concepts that no child has yet attempted, providing an empirically validated basis for distractor selection at cold start.
- Auditory scaffolding reliance falls between Tier 1 and Tier 2, consistent with the intended fading of support once a concept is recognised.
- The Graph Attention Network did not outperform the production heuristic or any classical baseline on any evaluated task, and was not promoted out of shadow mode.
- The distractor ranking task is confounded by the system's own exposure policy, which constructs the labels against which it would be measured. No offline claim about distractor quality is falsifiable without instrumented option-set logging.

7 DISCUSSION

7.1 Interpretation of Findings

The results divide cleanly into two groups: what the instructional design achieved, and what the graph intelligence strand did not.

The learning outcome is the clearer of the two. A mean within-concept gain of 0.358 score points across 69 repeated sequences, with 47 improving against 8 declining and a sign test p below 10 to the power of minus 7, establishes that repeated structured exposure within the module is associated with improved performance on the same concept. The within-subject design is what makes this defensible at 19 participants: each sequence serves as its own control, so the claim does not depend on comparison between children, and the number of children therefore does not limit the inference in the way it would for a between-subjects comparison.

The boundary of that claim must be stated plainly. It is a within-system learning effect. The design cannot establish transfer to contexts outside the module, cannot establish retention beyond the deployment period, and cannot attribute the gain specifically to the adaptive mechanisms as opposed to repeated exposure generally. Isolating the contribution of adaptivity would require a condition in which children received the same exposure without adaptive adjustment, which was neither run nor, given the participant numbers, runnable with adequate power.

The tier gate result requires particular care. Tier 2 shows a higher pass rate and a higher mean score than Tier 1, and the paired difference of 0.090 is statistically reliable. The tempting reading, that naming is easier than identifying, is wrong. The correct reading is that the gate works: children reach Tier 2 only after demonstrating recognition, so the Tier 2 population is a filtered population. The near-total disappearance of the lower tail at Tier 2 is the signature of selection, not of task difficulty, and the 98.95 per cent conditional pass rate confirms it. This is a design validation rather than a learning finding, and presenting it as the latter would be a misreading of the system's own architecture.

The similarity finding is the most substantial research result in this thesis, and its value lies in its specificity. The hypothesis was that children confuse concepts that resemble one another, and that computed resemblance could therefore substitute for observed confusion where none has been recorded. The data supports that hypothesis for visual resemblance with a medium effect size, and does not support it for phonetic resemblance. A system that weighted both features equally would be diluting a real signal with a null one.

The null phonetic result deserves interpretation rather than dismissal. The concept vocabulary is drawn from an early-primary curriculum in which words were selected for distinctness, so the phonetic similarity distribution is compressed, with a mean of 0.118 and most pairs scoring below 0.2. A compressed predictor has little room to discriminate. Furthermore, both tiers evaluated here are predominantly visual: Tier 1 presents images and Tier 2 presents written words. Neither requires the child to discriminate between spoken forms, which is where phonetic confusion would be expected to surface, and which is the domain of the separate pronunciation component. The honest conclusion is that phonetic similarity does not predict confusion in these two tiers, not that it is irrelevant to autistic vocabulary learning in general.

One observation complicates the purely visual account. The pair red with green appears among the most heavily weighted Tier 2 name confusions at weight 6, despite the written words sharing no visual resemblance and the colours themselves being maximally distinct perceptually. Their computed visual similarity is high because both assets are flat colour swatches of similar composition, which is an artefact of how the image encoder treats such images rather than a perceptual fact about the colours. This is a concrete limitation of the visual feature as currently computed, and it suggests that colour concepts require a different similarity treatment from object concepts.

The negative graph attention result is best understood as a data scale finding rather than an architecture finding. The three graph architectures were not merely uninformative on attempt outcome prediction but poorly calibrated, with higher log loss and Brier scores than the trivial baselines. That signature is characteristic of

overfitting, which is exactly what a parameterised model does when trained on several hundred examples drawn from 19 individuals. Against that, the classical baselines also failed: Rasch, the standard model for this task, achieved an ROC-AUC of 0.504. When the purpose-built classical model and the neural model both perform at chance, the limiting factor is the data rather than the model family.

The exposure policy confound is, in the judgement of this work, the most transferable finding of the research strand. A system whose fallback policy determines which distractors a child can select is a system that writes its own evaluation labels. Any offline metric computed on those labels measures the policy rather than the phenomenon. This failure mode is not specific to this system or to autism education: it applies to any adaptive learning platform that logs only what the learner chose and not what the learner was offered. The remedy, logging the option set on every scored attempt and introducing a small probability of randomised exposure, is inexpensive and should be considered a prerequisite for any claim about selection quality in this class of system.

7.2 Comparison with Literature, Baselines and Existing Solutions

Against the commercial products audited in Section 2.5, the module addresses each of the six gaps identified in Section 2.6. It provides a tiered mastery architecture that neither Otsimo nor Speech Blub offers. It implements a graduated relearn pathway that varies stimulus modality, against Otsimo's distractor removal and Speech Blub's identical repetition. It captures and applies sensory and typographic preferences that both lack. It positions the special education teacher as the operational user with authority over the pathway, against the parent-child model of both. It introduces attribute-level concept treatment at Tier 3. And it models concept relationships explicitly, which no reviewed system does.

The comparison with the research literature is more nuanced, and the contrast concerns evidence rather than features. Hayat et al. [9] report accuracy up to 100 per cent for teaching strategy suggestion, and Ganju et al. [10] report 98.85 per cent training and 97.70 per cent validation accuracy for handwriting-based cognitive

assessment. The present work reports an ROC-AUC of approximately 0.50 for attempt outcome prediction, which is unambiguously worse by that measure.

The difference is in what is being measured. Those studies evaluate retrospectively on assembled datasets, or, in Ganju et al.'s case, with three neurotypical children. The present work evaluates prospectively, under leave-one-student-out cross-validation, on data generated by autistic children in live classroom use, against the deployed heuristic the model would have to replace. A model that cannot beat a majority-class predictor under that protocol is a model that would not help a teacher, and reporting the figure is more useful than reporting a higher number obtained under conditions that would not survive deployment. Comparison against classical baselines is what makes the claim meaningful: implementing Rasch, Bayesian Knowledge Tracing, and the production heuristic before building the graph model is what converted a potential overclaim into an interpretable negative result.

Against the behavioural literature, the system's design choices find support. The structured, predictable, repetition-based architecture follows directly from the behavioural profile documented by Hariharan et al. [4]. The mastery criterion follows the ABA tradition reviewed by Wang and Wang [5] and the retention work of Fuller and Fienup [7], with the deliberate moderation discussed in Section 2.2. The dual-interface separation of teacher and child workspaces follows the finding of Akbar et al. [13] that task decomposition matters as much in the adult interface as in the child interface.

On attention as explanation, this work takes the conservative position. Attention weights were never surfaced to teachers as explanations of a child's cognition, which is consistent with the caution of Jain and Wallace and the qualified reply of Wiegreffe and Pinter. Since the model was not promoted, the question did not arise in deployment, but the design position was fixed before the result was known.

7.3 Research Contribution

The research contributions of this work are four.

The first is empirical evidence, from prospective deployment with autistic children in a Sri Lankan special education classroom, that a tier-gated, curriculum-aligned concept learning module produces reliable within-concept learning gains. The literature contains proposals for such systems; this work contributes evidence from one that was built, deployed, and measured.

The second is the similarity finding: that visual similarity between concepts predicts observed confusion with a medium effect size while phonetic similarity does not. This is an actionable result for the design of adaptive vocabulary systems in this population, and it is obtained at a unit of analysis, the concept pair, that makes it defensible despite a small participant group.

The third is the demonstration that this feature layer resolves the cold-start problem for 96.3 per cent of untouched curriculum concepts, which is the specific point at which per-item adaptive systems have nothing to offer.

The fourth is methodological: the identification and documentation of the exposure policy confound, and the accompanying negative result for graph attention at this data scale. A negative result reported with the analysis that explains it is a contribution, and in this case it is the contribution most likely to transfer to other work in adaptive learning.

7.4 Product Contribution

The product contribution is a working, deployed component of a larger platform, in use in a Sri Lankan special education school, comprising a three-tier concept learning module aligned to the national curriculum, a rule-based adaptive scaffolding engine with graduated relearn and bounded failure, a Graph Knowledge Base with append-only attempt history, a validated concept similarity feature layer, and a teacher-facing analytics and reporting surface integrated with three sibling modules.

Two engineering properties of that product are worth stating as contributions in their own right. The distractor selection chain degrades through three independent sources, so the module continues to function correctly with the graph service entirely unavailable. And the intelligence layer is isolated behind a hard timeout and a kill

switch, so that research components can be deployed in shadow mode without any possibility of affecting what a child experiences. These properties are what made it responsible to run a research model in a live classroom serving a vulnerable population.

7.5 Limitations and Threats to Validity

The limitations of this work are substantial and are stated here without mitigation.

Regarding external validity, 21 registered children from a single special education school cannot support generalisation to autistic learners broadly. Autism is defined by within-group heterogeneity, and a pooled result from one setting describes that setting. Curriculum coverage compounds this: fruits accounts for 64.2 per cent of recorded attempts, three categories have no data at all, and only 42 of 96 concepts have been attempted. Any claim about the module's behaviour on numbers, professionals, or house concepts would be unsupported.

Regarding internal validity, the observational design cannot separate the effect of the adaptive mechanisms from the effect of repeated exposure, and cannot rule out teacher effects, novelty effects, or maturation over the five-month period. The within-concept gain measure is computed only over sequences with more than one attempt, which is 69 of 299 sequences, and those sequences are not a random sample: a concept is reattempted precisely because the first attempt was unsatisfactory, so regression to the mean inflates the measured gain to an unknown degree. This is the most serious threat to the learning claim, and it is the reason the claim is stated as an association with repeated structured exposure rather than as a causal effect of the intervention.

Regarding construct validity, mastery is measured as the proportion correct within a three-trial block, which is a coarse instrument yielding only four possible values. Engagement is measured through tap counts and dwell times, which are behavioural proxies for attention rather than measures of it: a high replay count may indicate careful attention or may indicate difficulty, and the data cannot distinguish them.

Regarding the similarity analysis specifically, the unit of analysis is the concept pair, which pools evidence across children and therefore cannot speak to individual variation. Confused pairs are identified from observed confusions, which are themselves subject to the exposure policy confound: a pair never presented together cannot be observed as confused, so the non-confused group contains both genuinely non-confusable pairs and pairs that were simply never co-presented. This biases the comparison in an unknown direction, though the fact that the visual effect survives this contamination is modest evidence of its robustness. The red and green observation discussed in Section 7.1 further indicates that the visual feature behaves differently for colour concepts than for object concepts.

Regarding the graph intelligence evaluation, the distractor ranking results are not interpretable at all, for the reasons given in Section 6.2.1. The outcome prediction results are interpretable but are obtained at a data scale well below that at which graph neural networks are normally applied, so the negative result bounds what is achievable with this dataset rather than with the architecture in general.

Finally, no quantified usability instrument was administered, so the teacher validation reported in Section 6.3 is qualitative and indicative. A System Usability Scale score was specified in the research design but was not collected within the evaluation period.

7.6 Implications

For the design of adaptive learning systems in this domain, the clearest implication is that computed content features may deliver more value than learned models at realistic data scales. The similarity layer required no training data, is defined for the entire curriculum including untouched concepts, is interpretable, and produced a statistically reliable result. The graph attention model required substantial infrastructure and produced nothing usable. For a research group working with a classroom-sized participant group, feature engineering is likely the better investment.

For evaluation practice, the implication is that adaptive systems must log what the learner was offered and not only what the learner chose, and should introduce a small probability of randomised exposure. Without both, the system's own policy writes

the labels and offline evaluation measures the policy. This should be treated as a design requirement from the outset rather than a remediation.

For special education practice, the implication is that the teacher handover mechanism was the feature teachers valued most, which suggests that in this setting the value of an adaptive system lies as much in knowing when to stop and defer to a human as in deciding what to present next.

8 COMMERCIALISATION AND SUSTAINABILITY

8.1 Value Proposition and Target Users

The primary market comprises government and private special education institutions in Sri Lanka. Nationally there are approximately 27 dedicated special schools and 704 special education units. The principal of each institution serves as the purchasing decision-maker and maps directly to the system administrator role in the platform architecture. The direct end users are special education teachers and autistic children aged approximately five to twelve, who are non-paying beneficiaries.

Existing tools available to these institutions fall into two categories: generic global applications with no alignment to the Sri Lankan curriculum, and imported specialist resources that require trained therapists to administer. The value proposition addresses both shortcomings. The platform is the only structured, adaptive English language learning system built specifically around the Sri Lankan national syllabus and designed to be operated by a classroom teacher without a specialist technology background. By reducing the per-child preparation burden on teachers while delivering personalised, curriculum-aligned instruction, the system approximates outcomes that would otherwise require dedicated one-to-one support.

The evaluation reported in Chapter 6 strengthens this proposition in one specific respect that matters commercially. The cold-start coverage result means the system behaves sensibly on concepts no child at a given institution has yet attempted, which is the condition every new deployment begins in. A system that requires months of accumulated data before it adapts usefully is a difficult sell to an institution evaluating it on a trial basis.

8.2 Market and Competitive Position

The competitive advantages of the module are curriculum alignment to the national syllabus, a structured three-tier architecture with mastery-gated progression, a teacher-centred design providing educators with analytics and interpretive feedback, sensory-friendly interface personalisation, and the graph-informed adaptive layer.

The honest competitive position regarding that last point is that the graph layer currently contributes through its similarity features and its serving role rather than through learned inference, and marketing it as artificial intelligence driven personalisation would overstate what Chapter 6 supports. The defensible claim is that the system adapts using an empirically validated model of which concepts children confuse, which is both true and distinctive: no audited competitor models concept relationships at all.

A secondary procurement pathway runs through the Ministry of Education. Should the system demonstrate measurable outcomes at pilot scale, Ministry-level adoption could facilitate national rollout through centralised procurement.

8.3 Business and Sustainability Model

The system operates on a software as a service subscription model, with an annual licence fee charged per institution rather than per student, which makes budgeting predictable for schools operating under fixed government allocations. A freemium pilot licence allows institutions to evaluate the platform before committing, which aligns with how procurement decisions are typically made in Sri Lankan public education, where trial evidence is often required prior to budget approval.

Three subscription tiers are proposed based on institutional size: up to 30 students for small special education units, 30 to 120 students for mid-scale schools and larger units, and 120 or more students or multi-campus institutions for large organisations with expanded administrative needs.

Development costs for the six-month prototype period totalled approximately Rs. 34,500, covering cloud hosting, database hosting, language model API usage, mobile development tooling, design resources, connectivity, and transport for school visits and usability sessions. Ongoing costs comprise server hosting, database storage, language model API usage, and periodic curriculum-aligned content updates. The cost structure scales with the number of institutions and students onboarded rather than with usage intensity, which makes per-institution pricing a reasonable fit.

One cost implication follows directly from the Chapter 6 result. Because the graph attention model was not promoted, the production deployment does not carry a machine learning serving stack, and the inference container remains small. The research dependencies are confined to an offline specification. This materially reduces hosting cost relative to the architecture that would have been required had the model shipped.

8.4 Scalability, Intellectual Property and Compliance

The architecture scales horizontally at the application tier, with the backend and Python services containerised and stateless behind a reverse proxy. The data tier uses managed PostgreSQL and managed Neo4j, both of which scale within their respective service tiers. The present deployment runs on a connection-constrained database tier appropriate to pilot scale, and a production rollout across multiple institutions would require a higher tier.

Intellectual property comprises several protectable elements. The software architecture and adaptive learning algorithms, including the rule-based scaffolding engine, the tier-based concept delivery mechanism, and the learning analytics framework, are protectable through software copyright. The interface design, including sensory-friendly themes and avatar-based feedback mechanisms, may be protectable under design rights. Original educational content developed for the platform, including concept images, animations, and learning activities, is protected under copyright ownership. Ownership may be retained by the research team or transferred to an institutional technology commercialisation office depending on university policy.

Compliance obligations centre on the handling of minors' disability-linked behavioural data. The present system transmits no personally identifiable information to external services and stores all data in access-controlled infrastructure. A commercial deployment would require three additions beyond the current research arrangements: a documented data retention and deletion policy, explicit consent covering secondary use of interaction data for system improvement, and a defined

procedure for handling data belonging to a child who withdraws. These are identified in Section 9.4 as prerequisites rather than enhancements.

8.5 Contribution to the Sustainable Development Goals

This work aligns with SDG 4 on Quality Education, which calls for inclusive, equitable, and disability-sensitive education for all children. By delivering curriculum-aligned personalised learning to autistic children within the Sri Lankan special education context, the system addresses a gap in quality English language instruction for a population frequently excluded from mainstream educational support.

It aligns with SDG 10 on Reduced Inequalities by reducing the educational disparity between autistic learners and their neurotypical peers, and by addressing the broader inequity between the educational technology resources available in high-income countries and those accessible to teachers and students in Sri Lanka.

It aligns with SDG 3 on Good Health and Well-Being through design decisions that treat the sensory and emotional well-being of autistic children as a primary constraint rather than a refinement: calming visual environments, predictable interface behaviour, non-punitive error handling, and the bounded failure mechanism that prevents a child from experiencing an unbounded sequence of failures.

It aligns with SDG 17 by demonstrating the application of information and communication technology as an enabling tool for inclusive education in a developing country context.

9 CONCLUSION AND FUTURE DIRECTIONS

9.1 Conclusion

This thesis presented the design, implementation, deployment, and evaluation of the Concept Learning Module of Auriva, together with the graph intelligence layer supporting it. The module teaches foundational English vocabulary to autistic children in Sri Lankan special education settings through a three-tier progression of visual familiarisation, single word recognition, and attribute introduction, governed by a rule-based scaffolding engine enforcing an Applied Behaviour Analysis mastery criterion, and supported by a Graph Knowledge Base modelling students, concepts, categories, and the confusion relationships between them.

The module was deployed in a live special education school and evaluated with 21 registered children over 27 active days, producing 441 scored attempts across 42 concepts. Tier 1 mastery was reached in 80.3 per cent of concept records and Tier 2 in 98.3 per cent, and performance within a concept improved reliably across repeated exposures by a mean of 0.358 score points. The tier gate behaved as designed, with 98.95 per cent of concepts passed at Tier 1 and subsequently attempted going on to pass at Tier 2.

The principal research finding concerns the concept similarity feature layer. Concept pairs that children actually confused carried significantly higher computed visual similarity than pairs they did not confuse, with a medium effect size, while phonetic similarity showed no reliable association. Because these features are defined for every concept including those no child has attempted, covering 96.3 per cent of the untouched curriculum, they resolve the cold-start problem at exactly the point where per-item adaptive systems have nothing to offer.

The Graph Attention Network developed over the knowledge graph did not outperform the deployed heuristic or any classical baseline on any evaluated task, and was consequently never promoted out of shadow mode. The analysis of why constitutes a contribution in its own right. The distractor ranking task was found to be confounded by the system's own exposure policy, which constructs the labels against which distractor quality would be measured, rendering offline claims

unfalsifiable without instrumented option-set logging. On attempt outcome prediction, neither the graph models nor the classical baselines, including the Rasch model that is standard for this task, achieved discriminative performance above chance, which indicates a data scale limit rather than an architecture failure.

The overall position of this work is therefore that a carefully structured, curriculum-aligned, teacher-mediated instructional design produced measurable learning outcomes with real autistic learners, that computed content features delivered a validated and actionable research finding, and that learned graph inference did not justify itself at this data scale. Reporting all three honestly is, in the judgement of this work, more useful to the field than reporting only the first two.

9.2 Achievement of Objectives

Table 9.1: Achievement of the specific objectives stated in Section 1.5

Objective	Status	Evidence
Three-tier concept learning structure	Achieved	Deployed; 247 tier records across 41 concepts (Section 6.1.2)
Graph Knowledge Base with static and dynamic layers	Achieved	570 nodes, 2,663 relationships; append-only history added (Section 4.4.3)
Rule-based adaptive scaffolding engine	Achieved	All 14 adaptive scenarios verified (Table 5.1)
Concept similarity feature layer and empirical test	Achieved	Visual similarity predicts confusion, $p = 0.00013$ (Section 6.1.4)
GAT evaluation against classical baselines	Achieved; negative result	Full LOSO evaluation; GAT did not outperform (Section 6.2)
Teacher-facing interpretive feedback and reporting	Achieved	Dashboard and report surface deployed (Section 5.2.5)
Deployment and evaluation with autistic children	Achieved	21 children, 27 active days, live classroom (Section 6.1.1)

Objective	Status	Evidence
Quantified usability measurement	Not achieved	SUS not administered within the evaluation period (Section 6.3)

9.3 Individual Research and Product Contributions

The research contributions are the empirical demonstration of within-concept learning gain from a tier-gated curriculum-aligned module deployed with autistic children; the finding that visual similarity predicts observed concept confusion with a medium effect size while phonetic similarity does not; the demonstration that this feature layer resolves cold start for 96.3 per cent of untouched curriculum concepts; and the identification and documentation of the exposure policy confound together with an honest negative result for graph attention at this data scale.

The product contributions are a deployed three-tier concept learning module aligned to the Sri Lankan national curriculum; a rule-based adaptive scaffolding engine with graduated modality-varying relearn and bounded failure terminating in teacher handover; a Graph Knowledge Base extended with append-only attempt history making it temporal and reproducible; a validated concept similarity feature layer; a teacher-facing analytics surface integrated with three sibling modules; and a fault-tolerant architecture in which the intelligence layer is isolated behind a hard timeout and a kill switch, so that research components can run in a live classroom without any possibility of affecting what a child experiences.

9.4 Future Directions

The following work is required, ordered by priority.

- Instrument option-set logging on all scored attempt types, not only activity attempts, and introduce a small probability of randomised distractor exposure. Until both are in place, no claim about distractor selection quality is falsifiable. This is inexpensive, requires no schema change, and is a prerequisite for every subsequent item in this list.

9. Repeat the ranking evaluation once several weeks of instrumented data exist, and re-examine the go/no-go data gates. The present evaluation should be regarded as establishing the baseline and the method rather than the answer.
10. Deploy the visual similarity feature as a fourth source in the distractor selection chain, positioned ahead of the sequential fallback. This is the direct, immediate application of the finding in Section 6.1.4 and does not require the graph model.
11. Develop a separate similarity treatment for colour concepts, which the red and green observation in Section 7.1 shows are poorly served by an image-embedding similarity computed over flat colour swatches.
12. Broaden curriculum coverage beyond the fruits and animals categories, which together account for 85 per cent of recorded attempts. Three categories currently have no interaction data and no claim can be made about them.
13. Administer a quantified usability instrument with the teacher participant group, which was specified in the research design but not completed within the evaluation period.
14. Correct the 12 activity records carrying no difficulty level, and add a validation check preventing scores from being written without their difficulty context.
15. Establish a documented data retention and deletion policy, confirm consent coverage for secondary use of interaction data in model development, and define the procedure for handling data belonging to a child who withdraws, including the status of any model already trained on it.
16. Revisit the graph model only once the data gates are met, specifically at least 25 students with concept data, at least 2,000 graded attempts, at least 200 failure events, and at least five categories with meaningful data. At present none of these thresholds is met, and training past a prototype before they are would repeat the error this evaluation was designed to avoid.

REFERENCES

- [1] World Health Organization, "Autism Spectrum Disorder," World Health Organization, 17 September 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>. [Accessed 11 March 2026].
- [2] Y. M. Rohanachandra, D. M. A. Dahanayake, L. T. Rohanachandra and G. S. Wijetunge, "Knowledge About Diagnostic Features and Comorbidities of Childhood Autism Among Doctors in a Tertiary Care Hospital," *Sri Lanka Journal of Child Health*, vol. 46, no. 1, pp. 29-32, 2017.
- [3] Ada Derana, "9000 Children Identified as Autistic in Sri Lanka," Ada Derana, 11 April 2025. [Online]. Available: <https://www.adaderana.lk/news.php?nid=107358>. [Accessed 11 March 2026].
- [4] D. Hariharan, S. K. Banerjee, A. M. A. Jinnah, S. B. Banu and G. S. P. Ghantasala, "A Comparative Study Revealing the Behavioural Difference Between Autistic and Healthy Adults," in *International Conference on Communication and Signal Processing (ICCSP)*, Chennai, 2024.
- [5] C. Wang and H. Wang, "The growing challenge of autism spectrum disorder: a comprehensive review of etiology, diagnosis, and therapy in children," *All Life*, vol. 17, no. 1, 2024.
- [6] M. Granpeesheh, J. Tarbox and D. R. Dixon, "Applied behavior analytic interventions for children with autism: A description and review of treatment research," *Annals of Clinical Psychiatry*, vol. 21, pp. 162-173, 2014.
- [7] J. L. Fuller and D. M. Fienup, "A Preliminary Analysis of Mastery Criterion Level: Effects on Response Maintenance," *Behavior Analysis in Practice*, vol. 8, no. 1, pp. 1-8, 2018.
- [8] A. Vijayan, S. Janmasree, C. Keerthana and L. Baby Sylva, "A Framework for Intelligent Learning Assistant Platform Based on Cognitive Computing for Children with Autism Spectrum Disorder," in *International CET Conference on Control, Communication, and Computing (IC4)*, 2018.
- [9] A. Hayat, T. Islam and M. Rahman, "Personalized Teaching Strategy Suggestion for Toddlers with ASD: A Machine Learning and Deep Learning Approach," in *27th International Conference on Computer and Information Technology (ICCIT)*, 2024.
- [10] G. Ganju, A. Kumar and M. Uma, "Design And Development of Cognitive Skills Enhancement for Neurodivergent Children using CNN," in *International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 2024.
- [11] P. V. Kumar, D. Balamurugan, K. Suriya and R. Reshma, "An Improved Learning Experience for English Language Using Big Data Analytics and AAC," in *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2023.

- [12] L. Hiryanto, M. Angelina, Jonathan, Hersinta, O. D. Hutagaol and Tony, "UI/UX Prototype of Visually Interactive Communication and Reading Aid for Autistic Children with Speech Disability," in Eighth International Conference on Informatics and Computing (ICIC), 2023.
- [13] G. S. Akbar, E. R. Kaburuan and V. Effendy, "User Interface (UI) Design of Scheduling Activity Apps for Autistic Children," in International Conference on Orange Technologies (ICOT), 2017.
- [14] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, "Graph Attention Networks," in International Conference on Learning Representations (ICLR), 2018.
- [15] S. Brody, U. Alon and E. Yahav, "How Attentive are Graph Attention Networks?," in International Conference on Learning Representations (ICLR), 2022.
- [16] S. Jain and B. C. Wallace, "Attention is not Explanation," in Proceedings of NAACL-HLT, Minneapolis, 2019.
- [17] S. Wiegreffe and Y. Pinter, "Attention is not not Explanation," in Proceedings of EMNLP-IJCNLP, Hong Kong, 2019.
- [18] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253-278, 1994.
- [19] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danish Institute for Educational Research, 1960.
- [20] National Institute of Education, *English Language Teacher's Guide: Grade 1 and Grade 2*, Maharagama: National Institute of Education, Sri Lanka.

GLOSSARY

Block. A group of three scored trials within a tier, over which the mastery criterion is evaluated.

Cold start. The condition in which a learner or concept has no interaction history, so that per-item adaptive methods have no basis for a decision.

Concept. A single vocabulary item taught by the module, such as apple or elephant, carrying image, audio, and attribute assets.

Confusion edge. A weighted relationship in the knowledge graph recording that one concept was selected when another was correct.

Exposure policy. The rule determining which options a learner is shown; where it also determines which errors can be observed, it confounds evaluation.

Graph Knowledge Base. The Neo4j graph holding students, concepts, categories, attempts, and the typed relationships between them.

Learnt. The state of a concept for a given child once both Tier 1 and Tier 2 have been passed.

Mastery criterion. The requirement of two correct responses out of three within a block before a tier may be passed.

Proxy user. The special education teacher, who operates the system on behalf of the child during a learning session.

Relearn pathway. The three-level remediation sequence triggered on failure to meet the mastery criterion, varying stimulus modality at each level.

Shadow mode. Deployment in which a model is called and its output logged, but the production decision is taken by the existing heuristic.

Tier. One of the three instructional stages for a concept: visual familiarisation, single word recognition, or attribute introduction.

APPENDICES

Appendix A: Research Instruments and Extended Results

A.1 Knowledge Graph Snapshot Composition

The snapshot analysed in Chapter 6 was read from the production Neo4j instance on 11 October 2026 using a read-only session, and comprises 570 nodes and 2,663 relationships.

Table A.1: Knowledge graph node and relationship composition

Element	Type	Count
Node	Attempt	444
Node	Concept	96
Node	Student	21
Node	Category	9
Relationship	PHONETIC_SIMILARITY	534
Relationship	VISUAL_SIMILARITY	462
Relationship	ATTEMPTED	444
Relationship	ON	444
Relationship	T1_SCORE	128
Relationship	T1_ENGAGEMENT	124
Relationship	T2_SCORE	121
Relationship	T2_ENGAGEMENT	114
Relationship	HAS_CONCEPT	93
Relationship	T2_NAME_CONFUSION	59
Relationship	CONFUSION	50
Relationship	ACTIVITY_SCORE	48
Relationship	ACTIVITY_FORMAT_CONFUSION	29
Relationship	ACTIVITY_CONFUSION	13

A.2 Most Strongly Weighted Observed Confusions

Table A.2: Highest weighted confusion pairs by tier

Tier	Concept presented	Concept selected	Weight
Tier 1 identification	apple	cherry	21
Tier 1 identification	apple	banana	20

Tier	Concept presented	Concept selected	Weight
Tier 1 identification	grapes	mango	15
Tier 1 identification	banana	grapes	9
Tier 1 identification	apple	mango	5
Tier 1 identification	mango	papaya	5
Tier 2 name recognition	apple	banana	7
Tier 2 name recognition	apple	cherry	7
Tier 2 name recognition	grapes	mango	7
Tier 2 name recognition	orange	papaya	7
Tier 2 name recognition	red	green	6
Tier 2 name recognition	passion fruit	watermelon	5

A.3 Dataset Versions Used

Four dataset versions were produced during the research period. Version v1 contained 732 attempts, 164 confusions, and 258 outcomes from 11 students, and was used for the initial baseline evaluation that identified the exposure policy confound. Version v4, used for the graph attention evaluation reported in Section 6.2, contains 864 attempts, 310 outcomes, 205 confusions, 84 engagement records, and 329 snapshot rows, read from 2,350 interaction events across 12 students with activity. The descriptive and similarity analyses in Section 6.1 operate on the 11 October 2026 graph snapshot described in A.1, which supersedes v4.

A.4 Reproduction

The dataset builder, the baseline implementations, and the evaluation harness are maintained in the project repository under the research directory of the graph service, separate from the serving code and with separate dependency specifications. Each dataset version carries a manifest recording row counts and a content hash per table, so that a result can be tied to the exact data that produced it.

Appendix B: Academic Integrity and AI Use Disclosure

This appendix discloses the use of generative AI and automated tools in the preparation of this thesis and the system it describes, in accordance with module policy.

Generative AI was used as a development aid during implementation, for code scaffolding, debugging assistance, and documentation drafting, in the same manner as other development tooling. All code was reviewed, tested, and integrated by the candidate, and the candidate is responsible for its correctness.

A large language model is a functional component of the delivered system, used to generate teacher-facing interpretive summaries from anonymised session data. This use is described in Sections 4.1, 5.2.5, and 3.6, and is subject to the non-diagnostic constraint documented there. No personally identifiable child information is transmitted to that service.

Generative AI was used in preparing this document, for drafting assistance, structural organisation, and language editing. All quantitative results reported in Chapter 6 were computed from the project's own data by analysis code written for that purpose, and were verified against the source snapshot. All interpretations, conclusions, and limitations are the candidate's own.

The candidate remains responsible for the accuracy, originality, citations, code, data, analyses, and conclusions presented in this thesis.