

AURIVA: Adaptive English Language Learning Platform for Autistic Children in Sri Lanka

INDIVIDUAL THESIS

Maryse Christina Gunawardena

(IT22166838)

Individual component: Adaptive Dialogue Learning and Progression
Tracker

Group ID: R26-IT-006

Report submitted in partial fulfilment of the requirements for the

Bachelor of Science (Honours) in Information Technology

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. I remain responsible for the accuracy, originality, citations, code, data, analyses, and conclusions in this thesis.

Student name: Maryse Christina Gunawardena

Student ID: IT22166838

Signature: _____ Date: _____

ABSTRACT

Autistic children in Sri Lanka who learn English as a second language have few tools that follow the national curriculum, support children who do not speak, or allow for the way Sinhala shapes their English pronunciation. Existing systems teach vocabulary but rarely teach when to use a word, treat transfer errors and echolalia as failures, and give teachers little explanation of a child's progress. This research designed, built and evaluated the Adaptive Dialogue Learning and Progression Tracker, one component of the AURIVA learning application. Following Design Science Research, the component teaches 30 social words through three phases (familiarisation, production and contextual awareness) and short personalized sentences in a second level, with a non-verbal pathway throughout. It makes three research contributions: a phoneme-aware speech assessment that gives partial credit to Sinhala transfer errors; a latency-based rule that detects probable echolalia without penalising the child; and an explainable two-tier model that predicts each child's learning pace from their first session with a word. The component passed 114 unit tests and is being evaluated in a supervised field study with 10 children and about six teachers at three schools in the Colombo district. On 87 real first-session records, evaluated by leaving out one child at a time, a Random Forest predicted learning pace clearly better than a literature-grounded formula on the spoken attempts it handles (balanced accuracy 0.695 against 0.498). The field results for speech assessment, echolalia detection and usability are pending. The research contributes a deployed, curriculum-aligned learning component and evidence on how learning models for small special education populations should be trained, audited and evaluated.

Keywords: autism, English as a second language, speech assessment, echolalia, learning-trajectory prediction

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors for the continuous guidance and support provided throughout this research. I extend my appreciation to the staff and special educators at the partner special education school for their invaluable insights and cooperation during the requirement gathering process. I also thank my project group members for their collaboration throughout the proposal. Finally, I am grateful to my family for their encouragement and support throughout this journey.

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.....	6
1.6 Research Questions or Hypotheses	7
1.7 Scope, Limitations and Assumptions.....	7
1.8 Individual Contribution	8
1.9 Thesis Structure.....	8
2 LITERATURE REVIEW AND RELATED WORK	10
2.1 Review Method and Scope.....	10
2.2 Domain and Theoretical Foundations	10
2.2.1 Pragmatic language in autism	10
2.2.2 Autistic learners of English as a second language	11
2.2.3 Echolalia.....	11
2.2.4 First-language transfer from Sinhala.....	12
2.2.5 Staged learning and prompting	12
2.3 Related Research.....	12
2.3.1 Technology-assisted vocabulary and social communication interventions	12
2.3.2 Virtual agents, social robots and tutors	13
2.3.3 LLM-based dialogue systems	13
2.3.4 Speech recognition and assessment	14

2.3.5 Adaptive learning, progress prediction and explainability	15
2.4 Existing Systems, Products and Techniques	15
2.5 Gap Synthesis and Conceptual Framework	16
3 RESEARCH METHODOLOGY	18
3.1 Research Design.....	18
3.2 Data, Participants or Study Materials.....	21
3.2.1 Curriculum and learning content.....	21
3.2.2 Media and audio materials	22
3.2.3 Pilot interaction data	23
3.2.4 Real training data	24
3.2.5 Synthetic data (earlier design iteration)	24
3.2.6 Field study participants	25
3.3 Experimental or Study Procedure	25
3.3.1 Level 1 learning flow and mastery rules	25
3.3.2 RC1: Phoneme-aware speech assessment.....	29
3.3.3 RC3: Echolalia detection through response latency.....	30
3.3.4 RC2: Learning-trajectory prediction.....	31
3.3.5 Level 2 sentence construction	36
3.3.6 Software verification.....	36
3.3.7 Field study procedure.....	36
3.4 Measures and Evaluation Metrics	39
3.5 Data Analysis	40
3.6 Ethics, Privacy, Security and Responsible AI.....	41
3.7 Reproducibility and Research Artefacts.....	42
3.8 Commercialisation Aspect of the Product.....	44
3.8.1 Value proposition and target users	44

3.8.2 Market and competitive position.....	44
3.8.3 Business and sustainability model	45
3.8.4 Scalability, intellectual property and compliance	46
3.8.5 Contribution to the Sustainable Development Goals	46
4 PRODUCT AND SOLUTION DESIGN.....	47
4.1 Requirements.....	47
4.1.1 Stakeholders	47
4.1.2 Requirements from classroom observation.....	47
4.1.3 Functional requirements.....	48
4.1.4 Non-functional requirements	50
4.2 Component Architecture within the Complete Solution.....	51
4.2.1 Place in the AURIVA application.....	51
4.2.2 Internal architecture	52
4.2.3 Main data flows.....	53
4.3 Design Alternatives and Decisions	54
4.4 Detailed Design.....	56
4.4.1 Level 1 learning flow	56
4.4.2 Production phase and prompt hierarchy.....	58
4.4.3 Mastery rules and progression	59
4.4.4 RC1: Phoneme-aware speech assessment.....	61
4.4.5 RC3: Echolalia detection.....	62
4.4.6 RC2: Learning-trajectory prediction and explanation.....	63
4.4.7 Teacher reports.....	66
4.4.8 Level 2 sentence construction	67
4.4.9 Data design.....	70
4.4.10 Child-safety design principles.....	73

4.5 Standards and Professional Practice.....	73
5 IMPLEMENTATION, INTEGRATION AND TESTING.....	75
5.1 Development Environment and Technology Stack	75
5.2 Component Implementation.....	75
5.3 Integration with the Other Components.....	78
5.4 Verification and Software Testing	79
5.5 Deployment and Operation	80
6 RESULTS AND VALIDATION	82
6.1 Research Results	82
6.1.1 RC2: Algorithm comparison on synthetic data.....	82
6.1.2 RC2: Calibration audit	83
6.1.3 RC2: Synthetic-trained model on the real pilot set.....	83
6.1.4 Label audit.....	84
6.1.5 RC2: Final model trained on real data	84
6.1.6 RC1: Speech assessment agreement	86
6.1.7 RC3: Echolalia detection.....	86
6.2 Product and Technical Results	87
6.3 User or Stakeholder Validation.....	87
6.3.1 Teacher task usability.....	87
6.3.2 Teacher feedback on the component's judgements	88
6.3.3 Child engagement and observations.....	88
6.4 Integration Results	88
6.5 Summary of Findings.....	88
7 DISCUSSION	90
7.1 Interpretation of Findings.....	90
7.2 Comparison with Literature Baselines and Existing Solutions.....	92

7.3 Research Contribution.....	93
7.4 Product Contribution.....	94
7.5 Limitations and Threats to Validity	94
7.6 Implications.....	97
8 CONCLUSION AND FUTURE DIRECTIONS	98
8.1 Conclusion	98
8.2 Achievement of Objectives	99
8.3 Individual Research and Product Contributions	99
8.4 Future Directions.....	100
REFERENCES.....	102
GLOSSARY.....	107
APPENDICES	108
Appendix A: Research Instruments or Extended Results	108
Appendix B: Academic Integrity and AI Use Disclosure.....	108

LIST OF FIGURES

Figure 2.1: Conceptual framework	17
Figure 3.1: Design Science Research cycles applied to this research.....	21
Figure 3.2: Level 1 flow for one word across the three phases.....	28
Figure 3.3: RC2 trajectory prediction pipeline	35
Figure 4.1: Integrated AURIVA architecture and the boundary of this component..	51
Figure 4.2: Internal layered architecture of the component	53
Figure 4.3: Sequence of a Phase 2 spoken attempt	54
Figure 4.4: Level 1 screens for one word ("Thank You").....	58
Figure 4.5: Word status states and transitions.....	61
Figure 4.6: Trajectory prediction and explanation flow.....	65
Figure 4.7: Level 2 flow.....	69
Figure 4.8: Level 2 screens (Self-Introduction)	70
Figure 4.9: Entity-relationship diagram of the component's tables	72

LIST OF TABLES

Table 1.1: Comparison of related research	4
Table 1.2: Mapping of proposal objectives to final objectives	6
Table 2.1: Existing products and systems	15
Table 2.2: Capability coverage in the reviewed research	16
Table 3.1: Level 1 word set.....	21
Table 3.2: Mapping of mastery outcomes to trajectory labels.....	24
Table 3.3: RC1 scoring rules.....	29
Table 3.4: Terms of the Tier 1 baseline formula.....	31
Table 3.5: Evaluation measures by component.....	39
Table 3.6: Customer personas	44
Table 3.7: Revenue model.....	45
Table 4.1: Requirements from the partner school and classroom observation	47
Table 4.2: Functional requirements and their final status	48
Table 4.3: Non-functional requirements and design response	50
Table 4.4: Main design decisions.....	55
Table 4.5: Sinhala L1 transfer-error taxonomy used by RC1	62
Table 4.6: Trajectory model features	63
Table 4.7: Traceability of Level 2 activities to the Grade 1–2 activity book.....	67
Table 4.8: Level 2 target sentence patterns.....	68
Table 4.9: Database tables owned by the component	71
Table 5.1: Technology stack	75
Table 5.2: Automated test suites	79
Table 5.3: Main defects found and their resolution	79

Table 6.1: Algorithm comparison on synthetic data (n = 400, leave-one-child-out cross-validation).....	82
Table 6.2: Synthetic-trained Tier 2 and Tier 1 on the real pilot set	83
Table 6.3: Tier 1, synthetic-trained and real-trained models on the real dataset (RF: Random Forest; the real-trained RF is evaluated leave-one-child-out)	85
Table 6.4: Product and technical results available for this draft	87
Table 6.5: Summary of findings by research question.....	88
Table 8.1: Achievement of objectives.....	99
Table B.1: AI tools used and how their outputs were checked	Error! Bookmark not defined.

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AAC	Augmentative and Alternative Communication
AI	Artificial Intelligence
API	Application Programming Interface
ASD	Autism Spectrum Disorder
CMU	Carnegie Mellon University
DSR	Design Science Research
EMA	Exponential Moving Average
ER	Entity–Relationship
ESL	English as a Second Language
FR	Functional Requirement
L1	First Language
LLM	Large Language Model
NFR	Non-Functional Requirement
RBF	Radial Basis Function
RC	Research Contribution
RQ	Research Question
SEQ	Single Ease Question
SHAP	SHapley Additive exPlanations
SMOTENC	Synthetic Minority Over-sampling Technique for Nominal and Continuous features
SOV	Subject–Object–Verb
SSML	Speech Synthesis Markup Language
STT	Speech-to-Text
SUS	System Usability Scale

Abbreviation	Meaning
SVO	Subject–Verb–Object
TTS	Text-to-Speech
UTC	Coordinated Universal Time

LIST OF APPENDICES

Appendix	Description	Page
Appendix A	Research Instruments or Extended Results	109
Appendix B	Academic Integrity and AI Use Disclosure	109

1 INTRODUCTION

1.1 Overall Project Context

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition marked by differences in social communication and interaction, together with restricted and repetitive patterns of behaviour. One of the most affected areas of language is pragmatic competence: the ability to use language in social situations, such as greeting someone, saying thank you, answering a question or introducing oneself.

The Global Burden of Disease Study 2021 estimated that about 61.8 million people worldwide are autistic, around 1 in 127 [1]. In the United States, prevalence among 8-year-old children rose to 1 in 31 in 2022 [2]. In Sri Lanka, an early study estimated a prevalence of 1.07% among children aged 18 to 24 months [3], and by 2025 more than 9,000 children had been formally identified as autistic, with many more thought to be unidentified [4]. There is no population-based screening programme, and few speech and language therapists work in the public school system.

English is a second language for most Sri Lankan children, and it is the language of higher education and much of employment. For autistic children, the communication differences of autism combine with the demands of learning a new language [5], [6], so functional spoken English is one of the hardest skills for them to develop in a normal classroom. Special education teachers are expected to teach the national English curriculum while meeting very different individual needs, and often rely on improvised materials or on mainstream applications that were not designed for autistic children.

AURIVA is an adaptive English learning platform for autistic children in Sri Lankan special education, aligned to the national primary English curriculum. It was developed by a group of four students as four components, integrated into one mobile application with a shared login, shared student profiles and a shared database:

- **Graph Neural Network-Enhanced Adaptive Concept Learning Module**, which teaches English vocabulary and attributes through visual familiarisation, word recognition and attribute introduction;

- **Motor-Informed Adaptive Handwriting Learning Module**, which teaches letter formation through tracing, adapted to each child's motor ability;
- **Phoneme-Aware Adaptive Pronunciation Support Module**, which teaches the sounds of English through images, audio and lip-sync guidance, with pronunciation scoring;
- **Adaptive Dialogue Learning and Progression Tracker**, the focus of this thesis, which teaches functional spoken dialogue words and short sentences.

The work supports Sustainable Development Goal 4 (Quality Education), which calls for inclusive education for children with disabilities, and Goal 10 (Reduced Inequalities). It falls within the Software Systems and Technologies research cluster.

1.2 Individual Component Context

The Adaptive Dialogue Learning and Progression Tracker teaches autistic children the words and short sentences they need for everyday spoken interaction in English. It has two levels.

Level 1, Dialogue Word Acquisition, teaches 30 words and phrases in three categories: Greetings (for example, "Hello", "Good morning" and "How are you?"), Magic Words (for example, "Thank you" and "Excuse me") and Abilities (action words such as "Clap", "Eat" and "Sleep"). Each word passes through three phases mapped to Bloom's taxonomy: Familiarisation, where the child sees and hears the word in context; Production, where the child says the word and the system assesses it; and Contextual Awareness, where the child chooses the situation in which the word fits.

Level 2, Sentence Construction, teaches the spoken topics in the Grade 3 speaking criteria: introducing oneself, describing a friend and describing a pet. The sentences are personalised from each child's own details, and every activity traces to a page of the national Grade 1-2 English activity book.

Both levels include a non-verbal pathway, so children who do not speak can take part and progress, and teachers receive reports that explain each child's progress. The component makes three research contributions (RC), each aimed at a problem the original design could not solve:

- **RC1, phoneme-aware speech assessment.** Spoken responses are scored by aligning their sounds with the target word. Errors typical of Sinhala speakers learning English are recognised and given partial credit, rather than being treated as either correct or failed.
- **RC2, learning-trajectory prediction.** A two-tier model predicts whether a child is progressing at a fast, typical or struggling pace on a word from the features of the child's first session with it. The prediction adjusts the difficulty of what is offered next and is explained to the teacher.
- **RC3, echolalia detection.** The system measures how quickly a child begins to speak after the model word, flags probable immediate echolalia (repeating what was just heard), and stops it from counting as mastery.

The component was designed through several documented iterations (Section 3.1). The proposal included a third level that used a large language model (LLM) to generate whole conversations from the child's mastered vocabulary. Test rounds with children showed that, although they could learn individual words, holding a conversation was too difficult at their current level, and that conversations generated afresh each time were likely to confuse or unsettle them because they were unpredictable. Level 3 was therefore removed, and the effort went into helping children become familiar and comfortable with basic social words and simple sentences first: the trajectory prediction (RC2), echolalia detection (RC3) and the speech assessment that supports them (RC1). RC1 replaced the fuzzy string matching planned in the proposal, and RC3 was added after the pilot.

1.3 Problem Statement

Autistic children in Sri Lankan special education who are learning English as a second language need to learn when and how to use everyday spoken words, not only how to recognise or repeat them. Existing tools do not meet this need, for four reasons:

1. **They teach words, not their use.** Technology-assisted vocabulary interventions mainly test whether a child can name or identify a word [7]. Digital tools for social communication target skills such as greetings and conversation, but the

review of them did not examine language learning itself [8], and pragmatic language interventions are mostly delivered by clinicians or caregivers, with little known about whether their effects generalise [9].

2. **Their speech assessment does not suit this population.** General speech recognition is less accurate for non-native child speakers [10]. Fuzzy string matching, the obvious way to tolerate transcription errors [11], also rewards the very first-language transfer errors a Sinhala-speaking child needs help with, such as "tank you" for "thank you" [12].
3. **They do not separate real production from echoing.** Echolalia occurs in up to 85% of autistic individuals who develop speech [13], so a repeat-after-me task can record a child as having learned a word they only echoed.
4. **They do not adapt to each child's pace in a way teachers can see and trust.** Adaptive tutoring can improve learning in young second-language learners [14], but tools for autistic learners mostly use fixed word lists and fixed progressions [7], and give teachers little evidence of why a child is moved on or held back.

These problems are sharper in Sri Lanka, where very few tools follow the national curriculum, there are few speech therapists in schools, and most children speak Sinhala as their first language, with its own patterns of transfer into English, including Subject–Object–Verb word order.

1.4 Research Gap

Table 1.1 compares the most relevant previous work against the five capabilities this component combines.

Table 1.1: Comparison of related research

Work	Teaches dialogue use in context	Speech assessment for non-native child speech	Detects echolalia	Adapts to pace, explained to teacher	Curriculum-aligned, partial credit
Jiang and Cohen [8] (review of 31 studies)	Partly (skills such as greetings)	No	No	No	No
Urrea et al. [7] (review of 13 studies)	No	No	No	No	No

Work	Teaches dialogue use in context	Speech assessment for non-native child speech	Detects echolalia	Adapts to pace, explained to teacher	Curriculum-aligned, partial credit
Lai et al. [15] (ASD-iLLM)	Dialogue, but text-based	No	No	No	No
Schodde et al. [14] (adaptive robot tutor, non-autistic children)	No	No	No	Adapts; no teacher reporting	No
Mishra et al. [16] (human-mediated LLM robot)	Social scenarios	No (human mediator)	No	No	No
Bosker [11]; Tran et al. [10] (non-autistic speakers)	No	Partly (transcript matching; accent correction)	No	No	No
This component	Yes (Phase 3)	Yes (RC1)	Yes (RC3)	Yes (RC2, with explanations)	Yes

The reviews show that digital tools for autistic children mostly target isolated skills. Jiang and Cohen reviewed 31 studies of digital technologies for social communication; the skills targeted included conversation and initiation, such as greetings, but the review did not examine language development or vocabulary learning. Urrea et al. reviewed 13 studies of vocabulary interventions and found mixed results (5 positive, 6 mixed, 1 negative and 1 with no effect), with most interventions relying on fixed word lists. LLM-based systems such as ASD-iLLM produce more natural dialogue but do not assess the child's speech. Adaptive tutors such as the robot of Schodde et al. adapt which tasks they present, but they need robot hardware, were tested with typically developing children, and take answers on a tablet rather than assessing speech.

No reviewed system combines contextual (pragmatic) word teaching, speech assessment that understands the learner's first-language errors, detection of echoed responses, pace prediction that teachers can inspect, and alignment with a national curriculum. That combination, for autistic Sinhala-speaking children, is the gap this research addresses.

1.5 Aim and Objectives

Aim. To design, build and evaluate an adaptive dialogue learning component that teaches autistic Sri Lankan primary school children functional English dialogue words and short sentences, assesses their speech in a way that tolerates first-language transfer and echolalia, adapts to each child's pace, and gives teachers explained evidence of progress.

Objectives.

1. **O1.** Implement a three-phase Level 1 learning flow (Familiarisation, Production, Contextual Awareness) with mastery rules and a non-verbal pathway.
2. **O2.** Implement curriculum-traced Level 2 sentence construction with element-based partial credit and a contrastive Sinhala-English word-order phase.
3. **O3 (RC1).** Develop phoneme-aware speech assessment that gives partial credit for Sinhala first-language transfer errors.
4. **O4 (RC3).** Detect probable immediate echolalia from response latency and prevent it from counting towards mastery.
5. **O5 (RC2).** Develop a two-tier learning-trajectory prediction model, trained on real interaction data, with explanations for teachers.
6. **O6.** Evaluate the component through software verification, computational evaluation and a field study with children and teachers.

The proposal's five objectives were revised as the design changed. Table 1.2 maps them to the final objectives.

Table 1.2: Mapping of proposal objectives to final objectives

Proposal objective	Final objective	Change
Three-phase dialogue word acquisition	O1	Kept
Vocabulary-constrained LLM dialogue generation	-	Removed with Level 3 after test rounds showed conversation was beyond the children's current level; replaced by O5
Multi-layer speech assessment with fuzzy matching	O3	Fuzzy matching kept only as a fallback; phoneme-aware assessment added
Mastery-gated progression and non-verbal fallback	O1, O5	Kept; adaptation extended by trajectory prediction

Proposal objective	Final objective	Change
Evaluation of accuracy, completion and validity	O6	Kept; targets used as reference points (Section 3.4)
-	O2	Sentence production moved from Level 3 to Level 2
-	O4	Added after the pilot

1.6 Research Questions or Hypotheses

- **RQ1 (RC1).** How closely does phoneme-aware assessment, with a Sinhala first-language error taxonomy, agree with a human rater's judgement of autistic children's spoken English words?
- **RQ2 (RC2).** Can features from a child's first session with a word predict whether they will progress at a fast, typical or struggling pace better than a literature-grounded formula, when the model is trained and tested on real data from children it has not seen?
- **RQ3 (RC3).** How accurately does response latency identify immediate echolalia in a repeat-after-model production task?
- **RQ4.** Can special education teachers use the component and its reports effectively, and do they find its judgements and explanations credible and useful?

1.7 Scope, Limitations and Assumptions

In scope: Level 1 (30 words in three categories) and Level 2 (three topics) of the dialogue component; the three research contributions; the teacher reports; the component's integration into the deployed AURIVA application; and a formative field study at three special education schools, including the partner school.

Out of scope: the other three AURIVA components; the shared platform features (login, student registration and the principal's views); LLM-generated conversation, removed with Level 3; and clinical diagnosis or therapy.

Limitations (discussed fully in Chapter 7):

- although testing covers three schools, the number of autistic children at each school is small, so the results are formative rather than generalisable;

- the real data available for training the trajectory model is limited, and the early pilot labels were found to be faulty;
- speech assessment depends on a commercial speech-to-text service that was not trained on autistic or Sinhala-accented child speech;
- field study results that depend on data collected after 11 October 2026 are pending in this draft.

Assumptions: a teacher is present at every session; children use a tablet with a working microphone and an internet connection; and the teacher confirms each child's personal details.

1.8 Individual Contribution

The researcher designed, built and evaluated the Adaptive Dialogue Learning and Progression Tracker: the Level 1 and Level 2 learning flows and screens; the component's backend services and database tables; the Python microservice for phoneme scoring, trajectory prediction and explanations; the synthetic data generator, the training and evaluation pipeline and the audits; the teacher reports; and the evaluation instruments and field study. The shared platform (login, student profiles, integration of the four components and deployment of the application) was provided by the group, with integration and deployment carried out by the group leader. Development was assisted by AI tools under a documented process, declared in Appendix B.

Contribution statement. This research contributes: a three-phase dialogue learning design that makes pragmatic word use an assessed learning step for autistic second-language learners; a phoneme-aware speech assessment method that recognises Sinhala first-language transfer errors and gives them partial credit; a latency-based method for detecting echolalia within a production task; and a two-tier, explainable learning-trajectory model, together with evidence on how such a model should be trained, audited and evaluated when real data is scarce.

1.9 Thesis Structure

Chapter 2 reviews the literature and existing systems and sets out the research gap. Chapter 3 describes the research methodology, including the commercialization

aspect of the product, Chapter 4 the product and solution design, and Chapter 5 the implementation, integration and testing. Chapter 6 reports the results, and Chapter 7 discusses them, including the limitations. Chapter 8 concludes and sets out future directions.

2 LITERATURE REVIEW AND RELATED WORK

2.1 Review Method and Scope

The review covered four areas that correspond to the component's design questions: (1) autism, pragmatic language, echolalia and English as a second language (ESL); (2) technology-assisted vocabulary and social communication interventions, including virtual agents, robots and large language model (LLM) dialogue systems; (3) automatic speech recognition and assessment for non-native and child speakers; and (4) adaptive learning, progress prediction and explainability, together with existing products. Sources were found in IEEE Xplore, the ACM Digital Library, ScienceDirect, PubMed and Google Scholar, using search terms that combined "autism" with "vocabulary", "pragmatic language", "dialogue", "ESL", "speech recognition", "echolalia", "adaptive learning" and "large language model". Systematic reviews were used to map each area, and individual studies were used for specific techniques. Recent work (2019 onwards) was preferred; older sources are cited only where they are foundational. Because research on autistic ESL learners in Sri Lanka is scarce, studies from comparable settings (Malaysia, the Gulf region, Vietnam and China) and studies of non-autistic second-language learners were included where their techniques transfer.

2.2 Domain and Theoretical Foundations

2.2.1 Pragmatic language in autism

Pragmatic language is the use of language in context. A meta-analysis of randomised controlled trials of pragmatic language interventions for autistic children (925 participants aged 21 months to 14 years) found a small but significant effect over control groups (Hedges' $g = 0.274$), with larger effects when caregivers were involved ($g = 0.760$) [9]. Whether these effects generalise and last remains "largely unknown", and most of the interventions were delivered by clinicians or caregivers. This points to a need for structured practice that a teacher can run in class. The component's third phase, in which the child chooses the situation where a word fits, makes this knowledge of context an explicit, assessed step.

2.2.2 Autistic learners of English as a second language

Two qualitative studies from Malaysia describe the barriers. In two urban autism centres, 45 learners with mild autism aged 7–10 found it hard to respond, to initiate interaction and to acquire vocabulary, while their 14 teachers found it hard to personalise materials and found repetition tiring [5]. A later study at one centre found that autistic children are visual learners, retain new words for a short time, and find words of more than three syllables difficult [17]. Parents and ESL teachers in a further study reported that visual aids, modelling and repetition are effective, and saw technology as "a supplement... not a replacement" for the teacher [6]. These findings support the component's visual, repetitive and teacher-assisted design and its short, graded word set.

2.2.3 Echolalia

Echolalia is the repetition of speech the child has heard. It occurs in up to 85% of autistic individuals who develop speech; immediate echolalia is usually defined as a repetition within two conversational turns of the original, and delayed echolalia as one produced later [13]. Other definitions describe immediate echolalia as following the model after "a brief latency" [18], and a systematic review of 15 intervention studies found no uniform definition at all [19]. None of these sources specifies a timing threshold, which is why RC3 operationalises "immediately after the model" as a measured response latency.

Recent work has changed how echolalia is understood. In eight Mandarin-speaking autistic boys, 61% of echoes served a function, and immediate echoes were used "to show engagement or to manage their own thought processes" [20]; delayed echoes in 47 children served interactive purposes or self-regulation [21]; and two autistic children understood the individual words inside their own delayed echoes [22]. Reviews conclude that "echolalia should be considered functional" and that reducing it without understanding its purpose is "contrary to principles of neurodiversity-affirming care" [19], and that echolalia "should not be stopped" [13]. Some interventions still aim to reduce it [23]. Repeating a prompted word can also be a step towards learning it [19]. The component therefore neither penalises an echoed response nor counts it as evidence that the word has been learned (RC3). Cohn et al.

also note that commands and questions requiring specific answers can increase immediate echolalia, which is relevant to a repeat-after-model task.

2.2.4 First-language transfer from Sinhala

Sinhala-speaking learners of English replace the dental fricatives in words such as "think" and "this", confuse /w/ with /v/, replace /f/ with /p/, confuse /z/ with /s/, struggle with word-final voiced consonants, and simplify consonant clusters, which are "extremely rare in Sinhalese speech", for example by inserting a vowel before "school" [12]. Sinhala also places the verb last (Subject–Object–Verb), whereas English places it in the middle (Subject–Verb–Object). Sri Lankan English has its own established norms, and judging it only against a native-speaker standard has been criticised [24], [25]. These patterns underlie the RC1 error taxonomy, its use of partial credit, and the Level 2 contrastive word-order screen.

2.2.5 Staged learning and prompting

The three Level 1 phases follow the revised Bloom's taxonomy, moving from remembering to applying to analysing [26]. The prompt hierarchy follows standard prompting practice in vocabulary teaching for autistic children: in a clinic-based single-case study of two autistic preschoolers, match-to-sample teaching with least-to-most and most-to-least prompting brought one child to 100% independent responding on 30 object words [27].

2.3 Related Research

2.3.1 Technology-assisted vocabulary and social communication interventions

Jiang and Cohen [8] reviewed 31 studies of digital technologies for social communication in autistic children and adolescents aged 2–17: virtual reality (13 studies), augmented reality (5), video modelling (4), touchscreen devices (3), combinations of these, and a game and a robot (1 each). The skills targeted were mainly responding and maintaining conversations (15 studies) and initiation (12). They concluded that "no single tool can address all social communication skills areas", noted that most studies were exploratory with small samples and rarely examined generalisation, and did not themselves examine language or vocabulary learning. Urrea et al. [7] reviewed 13 studies (2006–2022) of technology-assisted

vocabulary interventions on computers and tablets: five reported positive results, six mixed, one negative and one no difference, and instruction without technology often gave "more consistent and rapid generalization". They called for individualised assessment of what each child already knows. Together, these reviews show that technology can help but its effect is inconsistent; most interventions teach fixed word lists, and none assesses whether the child knows when to use a word.

2.3.2 Virtual agents, social robots and tutors

Cole et al. [28] used the Baldi animated talking head in vocabulary tutors with four stages (pretest, introduction, practice and test). In a pilot with six autistic children aged 3–13, five worked independently within one session and retained what they had learned over the summer, but speech recognition was disabled for these children "because of unacceptable accuracy". Milne et al. [29] built a virtual social tutor with conversation and bullying modules for 14 children with autism or Asperger syndrome; scores improved by 32% and 54% ($p < 0.05$), but speech input was again not viable, and reports for caregivers were left as future work. Schodde et al. [14] used a Nao robot with an adaptive knowledge-tracing model to choose tasks while 40 typically developing German children aged 4–7 learned nine English animal words; children who received explanations of the robot's adaptations learned more (3.73 vs 2.18 words, $d = 1.04$), and the robot re-engaged children in 76% of cases. Disengagement was, however, detected by a human operator, answers were given on a tablet, and the children were not autistic. Kim [30] proposed a game that teaches classroom conversation by linking emotions to situations, which has not yet been evaluated. These systems show that an animated agent can hold autistic children's attention and that adapting to the learner helps, but none assesses spoken production, all depend on special hardware or a human operator, and none reports progress to teachers in a structured way.

2.3.3 LLM-based dialogue systems

ASD-iLLM [15] fine-tuned LLMs on ASD-iLLM-8k, a Chinese dataset of 8,035 dialogue instances built from real clinical sessions with 74 children and 20 clinicians; clinicians rated its responses closest in style to real doctors, but it works with text only and "has not been tested in real clinical dialogue interventions". ASD-Chat [31]

combined speech recognition, GPT-3.5 and an avatar to run question-and-answer dialogues on five topics with 12 autistic children aged 4–11, and its responses were rated below those of a human interventionist (1.43 vs 1.90). Mishra et al. [16] used GPT-2 and BART to generate social scenarios for a NAO robot, with a human checking every item; it was evaluated with four experts and a child actor rather than autistic children, and its training data "was not built keeping in mind children with ASD". These systems show what LLMs can do, but also why the original Level 3 was removed: generated dialogue is unpredictable, has not been evaluated with autistic children in real settings, and needs a human to check it, which conflicts with the predictability autistic children need and with what the school test rounds showed (Section 1.2).

2.3.4 Speech recognition and assessment

Bosker [11] compared fuzzy string-matching metrics for scoring typed listener transcripts against human word-correct scores and found that the Token Sort Ratio correlated closely with them ($r = 0.922$ overall). It ignores word order and compares spelling rather than sounds, and it was tested with Dutch adult listeners. This is the method the proposal adopted and the reason RC1 was developed: a spelling-based metric cannot distinguish a transfer error such as "tank you" from a badly transcribed correct response, and it gives full credit to the very error the child needs to practise. Tran et al. [10] found that Google Speech-to-Text recognised only 62% of English sentences spoken by about 200 Vietnamese children aged 7–10 (for example, "Do you have a smartphone?" became "Do you have a microphone?"); a bag-of-words neural network that maps the transcript to the nearest curriculum sentence raised this to 80%. Their aim was to recover what the child meant, so the method hides pronunciation errors rather than assessing them.

Speech and language features have also been used to assess autism automatically. SPASHT [32] classified transcripts of 76 children as autistic or typically developing with 94.1% accuracy using semantic and pragmatic features, including an automatic echolalia feature that counts the child's immediate repetition of word pairs spoken by the therapist, and used SHapley Additive exPlanations (SHAP) to identify the most informative features. Its purpose is diagnosis rather than learning, and it detects

echoes from text rather than timing. Both Cole et al. and Milne et al. disabled speech recognition for autistic children. No reviewed work automatically assesses the pronunciation of autistic second-language learners, classifies their errors by first-language transfer, or separates echoed from genuine production during learning.

2.3.5 Adaptive learning, progress prediction and explainability

Predicting a learner's mastery from their responses has a long history in intelligent tutoring, starting with knowledge tracing [33], and predicting student performance is one of the main tasks of educational data mining and learning analytics [34]. Most such methods assume large amounts of data, which special education settings do not have. In autism-focused work, Farzana et al. [35] reviewed 60 studies of augmentative and alternative communication (AAC) from 2015–2020: mobile applications were the largest group (53.33%), only 6.67% used AI or machine learning, and the authors recommended combining intelligent analytics, personalisation and progress monitoring. Kumar et al. [36] proposed an eye-gaze AAC framework reporting 98.63% accuracy in simulation, with no participants described. Ranaweera et al. [37] built a game-based learning platform at SLIIT with 12 games across four domains for 20 autistic children aged 7–12, adjusting difficulty to each child's progress and reporting to parents and educators. When AI informs decisions about learners, teachers need to understand why; explainable AI is increasingly seen as necessary in education [38], and SHAP is a widely used method for explaining individual predictions [39]. None of the reviewed autism systems predicts a child's learning trajectory from interaction data or explains its adaptive decisions to teachers, which RC2 addresses.

2.4 Existing Systems, Products and Techniques

Table 2.1 compares existing products and research systems with the component.

Table 2.1: Existing products and systems

System	Type	Strengths	Limitations relative to this component
Proloquo2Go	Symbol-based AAC	Custom vocabulary, natural voices	iOS only; communication aid, not teaching; no speech assessment

System	Type	Strengths	Limitations relative to this component
Avaz	AI-assisted AAC	Predictive symbols, natural voices	Assistive, not instructional; no dialogue teaching
Speech Blubs	Speech practice app	Video modelling by children	Imitation-based; no context-of-use check; not curriculum-aligned
Mita	Language therapy app	Content adapts to the child	Receptive skills; no speech production or context-of-use check
TappyTots	Button-based AAC	Progress tracking	Basic needs only; no speech or adaptive teaching
AutiAid [40]	Learning app (UAE)	Memory, maths and alphabet games; teacher and parent access	No language or dialogue learning
VICARA [41]	AAC prototype (Indonesia)	Talking picture cards, bilingual, activity logs	Prototype; not tested with children
SLIIT platform [37]	Game-based learning (Sri Lanka)	Adaptive difficulty, progress reports	General skills; no English dialogue or speech assessment
This component	Adaptive dialogue learning	Context-of-use phase, RC1, RC2, RC3, curriculum-traced, explained reports	Small evaluation sample; Sinhala-speaking context only

2.5 Gap Synthesis and Conceptual Framework

Table 2.2 summarises how the reviewed research covers the six capabilities this component combines.

Table 2.2: Capability coverage in the reviewed research

Capability	Covered by	Gap
Assessing when a word is used	Partly: Milne et al. [29], Kim [30] and Mishra et al. [16] (social situations)	No system assesses the context of use of individual dialogue words
Automatic assessment of the child's speech	Partly: Tran et al. [10] (non-autistic children; corrects rather than assesses)	Speech recognition disabled for autistic children in earlier systems; no transfer-aware scoring
Echolalia detection	Partly: SPASHT [32] (from transcripts, for diagnosis)	No detection during learning, and none based on timing
Adapting to the individual child	Schodde et al. [14]; Ranaweera et al. [37]	No trajectory prediction from interaction data
Explained progress for teachers	Partly: AutiAid [40]; Ranaweera et al. [37]	No explanation of why the system adapts

Capability	Covered by	Gap
National-curriculum alignment	Partly: Tran et al. [10] (Vietnamese curriculum sentences)	No Sri Lankan curriculum-aligned dialogue tool

The gap is the combination: no reviewed system teaches autistic second-language learners when to use everyday words, assesses their speech with awareness of first-language transfer, separates echoed from genuine production without penalising it, predicts and explains each child's pace, and follows a national curriculum. Figure 2.1 shows the conceptual framework that links the literature to the design. Pragmatic language research and Bloom's taxonomy shape the three phases; the ESL and autism studies shape the visual, repetitive, teacher-assisted design; first-language transfer research shapes RC1; echolalia research shapes RC3; and adaptive tutoring and explainable AI shape RC2 and the teacher reports.

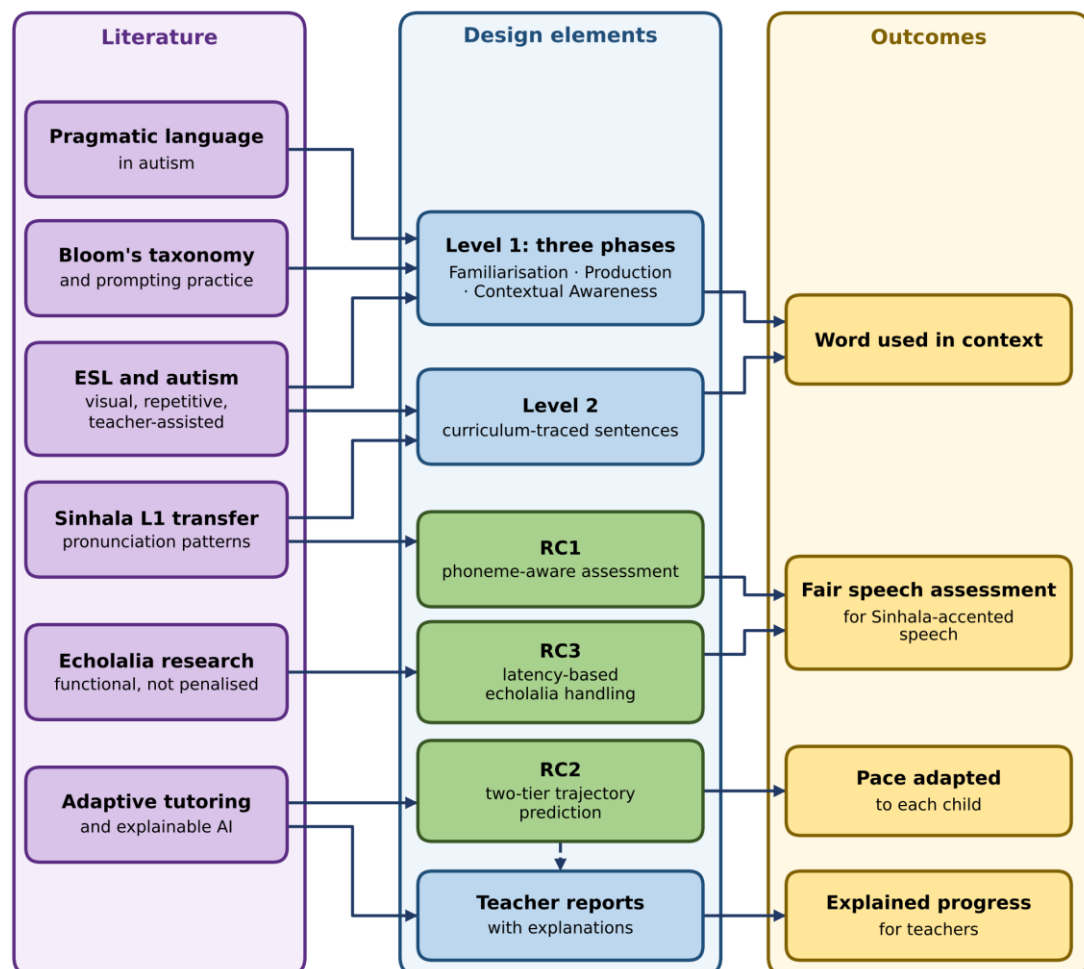


Figure 2.1: Conceptual framework

3 RESEARCH METHODOLOGY

This chapter describes how the Adaptive Dialogue Learning and Progression Tracker was designed, built and evaluated. It covers the research design; the curriculum, media and data; the procedure used to build the learning flow and each research contribution (RC); the measures and analysis used to evaluate them; and the ethical safeguards and the steps taken to make the work reproducible; and the commercialisation aspect of the product.

3.1 Research Design

This research follows the Design Science Research (DSR) methodology [42], [43]. DSR fits applied computing research whose main output is a working artefact built to solve a defined real-world problem. Here, the artefact is a mobile learning module that teaches functional English dialogue words and short sentences to autistic Sri Lankan primary school children, adapts to each child's progress, and gives teachers evidence they can act on. Figure 3.1 summarises how the three DSR cycles were applied.

Relevance cycle. The problem and requirements came from two sources. The first was the national curriculum: content was taken from the Sri Lankan Grade 1–2 English activity book. The second was the partner special education school, which asked for a dialogue-teaching module to complement the other three AURIVA components and described the Grade 3 spoken English criteria its pupils are assessed on:

- answering questions about name, age and home town;
- introducing oneself in at least four sentences;
- introducing a friend;
- speaking about a pet in at least three sentences.

The school advised that autistic children should be credited for attempts and partial answers, not only for complete ones. A school visit on 2 April 2026 collected audio samples and feedback, and later classroom observations directly shaped several design decisions: the non-verbal pathway, adaptive wait times for children who do

not speak, protection against rapid double-taps on the record button, extra familiarisation screens, and assessments with no failure state.

Design cycle. The module was developed iteratively, and its final form is the result of several documented design iterations:

- The original proposal included a third level, Conversation Builder, which generated whole conversations with a large language model (LLM). Test rounds with children showed that, although they could learn individual words, holding a conversation was beyond their current level, and conversations generated afresh each time were likely to confuse or unsettle them because they were unpredictable. Level 3 was therefore removed, and the effort moved to helping children become familiar and comfortable with basic social words and simple sentences first, through Levels 1 and 2, trajectory prediction and echolalia detection.
- A fourth Level 1 category, Days of the Week, was removed because sequential recall of day names does not exercise pragmatic, context-dependent use of a phrase.
- Speech assessment by fuzzy string matching was replaced by phoneme-aware assessment (RC1).
- Echolalia detection was added (RC3).
- The trajectory prediction model (RC2) went through three iterations: a model trained on synthetic data; the same model with a calibration step, removed after an audit; and the final model, trained only on real interaction data (Section 3.3.4).

Each change was recorded as a numbered decision in a design decision log, so the reasoning behind every design choice can be traced.

Rigour cycle. Design decisions were grounded in published work on technology-assisted vocabulary and communication interventions for autistic children [7], [8], echolalia in autism [13], [19], automated assessment of speech transcripts [11], Sinhala-English phonological transfer [12], and Bloom's taxonomy as the framework for staged learning [26]. The rigour cycle is closed by evaluating the artefact against

defined measures, both computationally and with children and teachers at the partner school.

Evaluation strategy. Each part of the artefact is evaluated with the method best suited to it:

- **Software verification** checks that the system behaves as specified.
- **Computational evaluation** measures the trajectory prediction model using leave one child out cross validation on real data.
- **A formative field study** with children and teachers, using the deployed and fully integrated AURIVA application, evaluates speech assessment, echolalia detection, usability and stakeholder acceptance in the real setting.

The field study aims to show that the artefact works and is useful in its target context. It does not aim for statistically generalisable claims, which DSR evaluation does not require [42].

Development process. The artefact was built through a series of small development tasks that could each be verified on its own, 49 numbered tasks in total, plus a number of smaller ad hoc fixes recorded in the status file. Each task had a written specification with explicit acceptance criteria. The planning records were kept as files alongside the code: a status file, the numbered decision log, a roadmap, the task specifications, and review and audit reports. The project's history therefore does not depend on any single tool or session. Three working rules were followed:

1. **Database before interface.** Schema and backend changes were made and verified before any frontend work.
2. **One category at a time.** Changes affecting child safety or the learning flow, such as backward navigation, were piloted in one word category before being extended to the others.
3. **Stop on mismatch.** Whenever a specification did not match the existing code, work stopped and the specification was corrected before continuing.

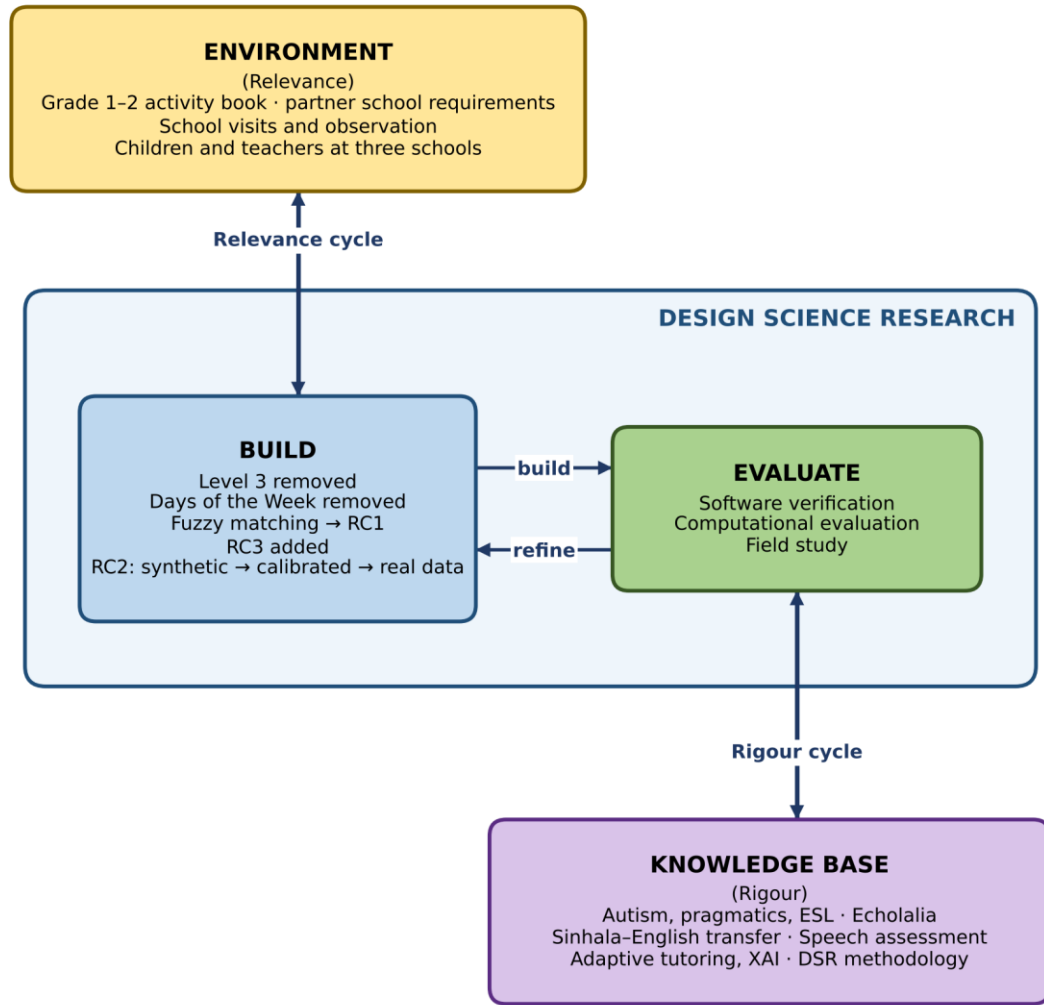


Figure 3.1: Design Science Research cycles applied to this research

3.2 Data, Participants or Study Materials

3.2.1 Curriculum and learning content

Level 1 (Dialogue Word Acquisition) teaches 30 words and phrases in three categories, listed in Table 3.1. Each word is graded by difficulty, and harder words are offered only after easier ones are mastered.

Table 3.1: Level 1 word set

Category	Difficulty 1	Difficulty 2	Difficulty 3
Greetings	Hello, Goodbye	Good morning, Good afternoon, Good night, Happy Birthday	How are you?, I'm fine, Happy New Year
Magic Words	Thank you	You're welcome, I'm sorry	Excuse me

Category	Difficulty 1	Difficulty 2	Difficulty 3
Abilities	Yes, No, Clap, Run, Walk, Jump, Talk, Dance, Sing	Brush, Wash, Eat, Drink, Write, Play, Sleep, Watch	-

Level 2 (Sentence Construction) teaches short spoken sentences in three topics: Self Introduction, Describing a Friend and Describing a Pet. Each activity can be traced to a page of the Grade 1-2 activity book. For example, the "Draw yourself" self-portrait activity follows page 1, the age picker follows page 8, and the pet picker is limited to the six animals on page 15 (cat, dog, cow, fish, parrot and rabbit). The topics also match the school's Grade 3 speaking criteria described in Section 3.1. The full traceability table is given in Chapter 4.

3.2.2 Media and audio materials

Each Level 1 word is taught through short scenario videos and illustrations set in culturally familiar Sri Lankan contexts; for example, "Happy New Year" uses Aluth Avurudu imagery. Videos were generated with Google Flow and Grok, and illustrations with Gemini and ChatGPT. Every generated asset was reviewed by the researcher and regenerated if it contained errors.

Audio comes from three sources:

- **Word and narration audio for Level 1** was recorded by a human speaker. Instructions are combined Sinhala and English narration, and the teacher can switch the Sinhala instructions off for each child. They are on by default, so support can be reduced as the child grows more confident.
- **Personalised Level 2 sentences** are generated with Azure text-to-speech (TTS) from the details in the child's profile. The first version used two voices: a British English voice for the English words and for eight internationally anglicised district names such as Colombo and Kandy, and a Sinhala voice, chosen by the child's gender, for the other 17 district names. At that time the only Sinhala voices available were one female and one male voice. When newer voices became available, the design was changed to a single voice for consistency: one high definition voice (en-IN-Aarti) now reads every sentence, and Sinhala names, meaning the 17 non-anglicised districts and the child's teacher confirmed name,

are marked as Sinhala within the same voice. The voice therefore no longer depends on the child's gender. Speech is slowed to 60% of the voice's normal rate, chosen by comparing samples at rates from 1.0 down to 0.6, because the normal speed sounded rushed for young learners. The single voice design was deployed during the field study; sessions before its deployment used the original two voice audio.

- **The child's own name** is given a suggested Sinhala script spelling, which a teacher must confirm before it is saved. The transliteration library was correct on 15 of 19 real names tested and produced confident errors on the rest.

3.2.3 Pilot interaction data

Real interaction data was collected at the partner school between 27 April and 2 July 2026. The first trajectory dataset extracted from it had 27 labelled rows from 6 children: 8 fast, 9 typical and 10 struggling. Of these, 15 rows came from the verbal pathway and 12 from the non-verbal pathway. Several data quality problems were found and corrected during development:

- The phoneme scoring service was not running during the early pilot sessions, so those attempts were scored by the fallback method and have no phoneme values.
- Per-scenario Phase 3 telemetry does not exist before 10 June 2026, and it was unreliable until 25 July 2026 because the pilot device had not been rebuilt with the current version.
- Response timestamps were missing for 42 Phase 2 attempts; for those attempts the echolalia flag kept its default value of false.
- One category produced duplicate rows for a phase. The cause was fixed, but some orphan rows still appear and must be removed during extraction.

A label audit on 6 October 2026 found a more serious problem. All ten "struggling" labels in the pilot data are wrong. Most came from two faults, both fixed on 6 August 2026: the consecutive-failure counter counted failures within a session as well as across sessions, so it could exceed the total number of sessions; and correct non-verbal sessions were scored as failures. Re-applying the corrected rules recovers none of these labels, because most of the affected words return to "in progress" and

one or two become "mastered". The fast and typical labels are logically sound but few: since the fix, about 14 rows from 6 children are trustworthy, and none of them is labelled struggling.

3.2.4 Real training data

The final trajectory model is trained only on real interaction data from two sources: pilot rows created after 6 August 2026 (rows from that day are excluded, because three label-related fixes were deployed on it), and data collected during the field study (Section 3.3.7). The field study records data in the same database as the pilot, so rows from both sources are extracted with the same script and filtered by creation date. Earlier pilot labels are not re-derived; they are kept only as evidence of the labelling fault reported in Chapter 7.

The extracted dataset contains 87 first-session rows, one per learner and word: 31 fast, 30 typical and 26 struggling. The accounts belong to children from the earlier pilot at the partner school and the children in the field study. Thirty rows, 17 of them struggling, come from the non-verbal pathway, and 57 are verbal. Phoneme accuracy is missing on 47 rows (always on non-verbal rows) and the Phase 2 latency on 23.

Labels come from the mastery rules, which are the sole authority for status (Section 3.3.1). Table 3.2 shows how each outcome is mapped to a label. Words still in progress at the end of data collection are excluded.

Table 3.2: Mapping of mastery outcomes to trajectory labels

Outcome	Label
Mastered in two sessions, the minimum possible	Fast
Mastered in more than two sessions	Typical
Struggling status under the corrected Rule 2	Struggling

3.2.5 Synthetic data (earlier design iteration)

Before enough real data existed, a generator produced a synthetic dataset of 400 rows (100 simulated children, four words each) with the same 13 features, using random seed 42. Five domain-randomised variants were then made by perturbing each generator coefficient by a random factor between 0.7 and 1.3, using seeds 42 to 46

[44]. This data was used to compare algorithms and to train the first version of the trajectory model. It is not used in the final model.

3.2.6 Field study participants

The field study involves three special education schools in the Colombo district, including the partner school. Because each school has only a small number of autistic children, testing was extended beyond the partner school. At the time of writing, 10 autistic children in the early primary grades have taken part across the three schools, including verbal and non-verbal or minimally verbal children, identified only by participant codes, together with about six teachers who observed and supervised the sessions and used the teacher-facing features.

3.3 Experimental or Study Procedure

3.3.1 Level 1 learning flow and mastery rules

Every Level 1 word moves through three phases, mapped to rising levels of Bloom's taxonomy [26]. Figure 3.2 shows the flow for one word.

Phase 1, Familiarisation (Remember). For Greetings and Magic Words, the child watches scenario videos and then sees the word on two screens: an animated, colourful version followed by a bold, static version. The child then places the word in a speech bubble by dragging it. The system counts exposures and records the ratio of exposures to the number required once, when the child passes the Phase 1 gate. For Abilities, a single tap-to-see interaction replaces the counted video exposures. Feedback is graded and never shows failure. If a child struggles twice in the same sitting, the video is replayed at most once before the activity moves on, which avoids repetitive loops.

Phase 2, Production (Apply). The avatar models the word and the child is invited to say it. The response is scored from 0 to 3 (Section 3.3.2) and checked for echolalia (Section 3.3.3).

- A score of 2 or more moves the child to Phase 3.
- A partial attempt (score 1) triggers a graded prompt hierarchy. On the second attempt, the difficult letters of the word are highlighted (for example, **T**hank you)

and the word is replayed at 60% speed. On the third attempt, a "Say it with me!" prompt is given with the word model.

- Three consecutive attempts with no recognisable speech (score 0) move the child to the non-verbal pathway, where the same goal is completed by tapping the correct picture.

Wait times also adapt across a day. Each consecutive refusal shortens the prompt waits (multipliers of 1.0, 0.7 and 0.5), and after three refusals the child goes straight to the non-verbal pathway for the rest of that day. These values are provisional defaults.

Phase 3, Contextual Awareness (Analyse). The child chooses the picture that shows a situation where the word fits, across scenarios A, B and C. If scenario A is answered wrongly, a checkpoint scenario is added. Phase 3 is passed when A, B and C are all correct, or when A is wrong but B, C and the checkpoint are correct. This phase tests whether the child knows when a word is used, not only whether they can say it.

Session outcome. A session passes when Phase 2 reaches its threshold and Phase 3 is passed. The Phase 2 threshold is a score of 2 or more on the verbal pathway and 1 on the non-verbal pathway, because non-verbal scores are binary.

Mastery rules. A word's status (not started, in progress, mastered or struggling) is set by fixed rules:

- **Rule 1 (mastery):** two passed sessions on different calendar days.
- **Rule 2 (struggling):** three consecutive failed sessions.
- **Rule 3 (reintroduction):** a struggling word returns once the easier words are mastered.
- **Rule 4 (Phase 1 pacing):** the number of Phase 1 exposures required for the child, four by default, is increased by one after three or more sessions without a pass, up to a maximum of six. The code also contains a branch that would reduce it after very fast mastery, but that branch cannot be reached in the current flow (Section 4.4.3).

- **Rule 5 (production probe):** every 14 days, words mastered through the non-verbal pathway are re-presented with a speaking prompt. If speech has emerged, the record of how the word was mastered changes from "non-verbal" to "mixed". This interval is provisional.

Other features. The system records whether each word was mastered verbally, non-verbally or through a mix of both. It also includes formative evaluations, in which the child matches words to pictures, unlocked when at least three words in a category are mastered and with no fail state; and an adult check that prevents a child from accidentally navigating backwards across a phase boundary. The trajectory model (Section 3.3.4) influences only which word is offered next and how long the Abilities familiarisation screens are shown. It never changes a word's mastery status.

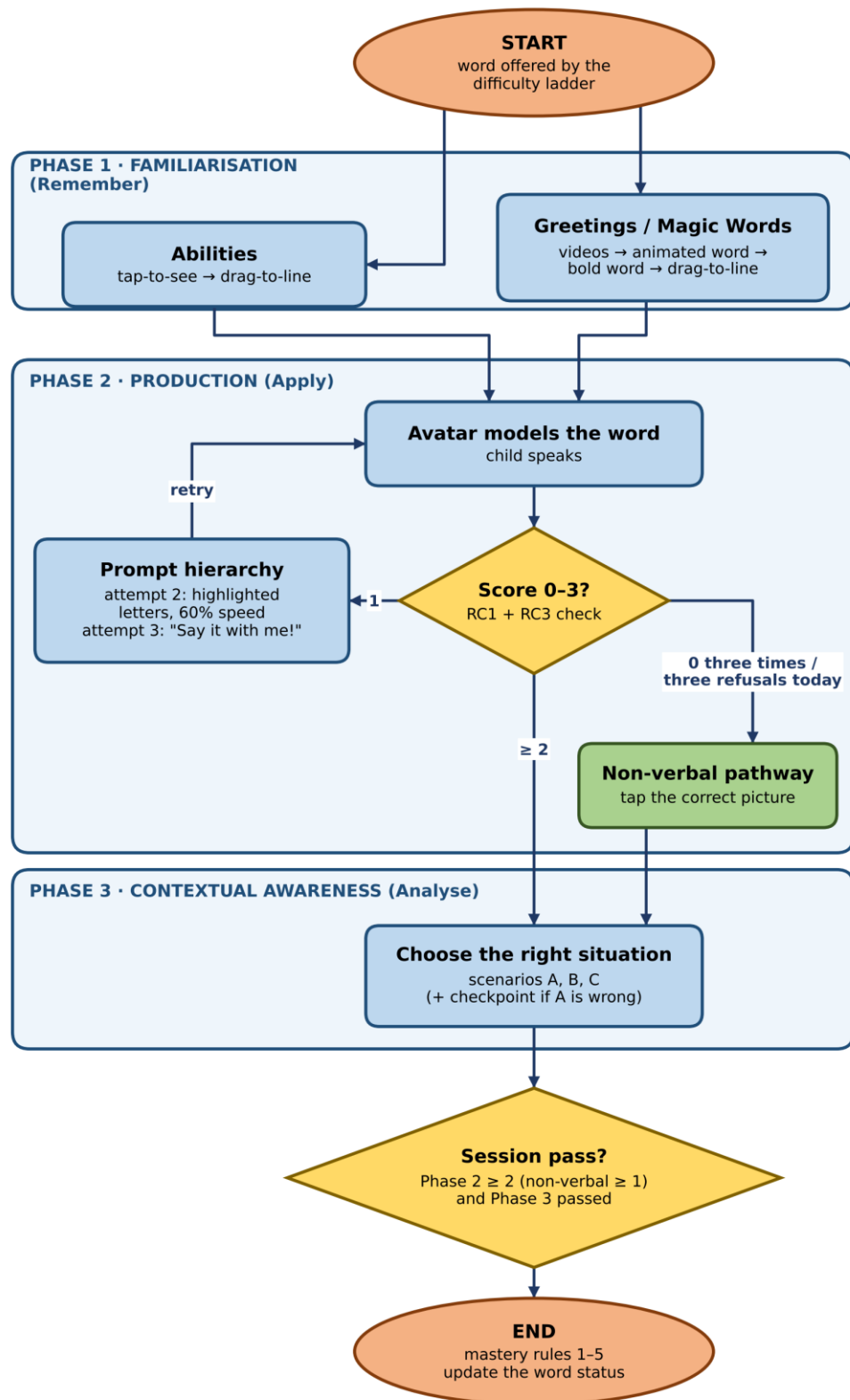


Figure 3.2: Level 1 flow for one word across the three phases

3.3.2 RC1: Phoneme-aware speech assessment

The proposal's speech assessment compared the speech-to-text transcript with the target word by fuzzy string matching [11]. That approach treats "tank you" as a near match for "thank you". For a child whose first language (L1) is Sinhala, however, replacing the "th" sound with "t" is a well-documented transfer error, because Sinhala has no dental fricative [12]. String matching therefore rewards exactly the error the child needs help with. RC1 works as follows:

1. The recorded audio is converted to an uncompressed WAV format.
2. Google Cloud Speech-to-Text transcribes it, using British English, the short-command model and word timings.
3. The transcript is first compared with the accepted forms of the target word. An exact match with an accepted full form scores 3, and a transcript containing an accepted key phrase scores 2. These attempts are not passed to phoneme scoring.
4. Any other transcript is sent to a separate Python service, which converts it and the target into phoneme sequences using the Carnegie Mellon University (CMU) Pronouncing Dictionary and aligns them.
5. Each difference is classified against a Sinhala L1 transfer-error taxonomy: TH-fronting and TH-stopping, DH-stopping, final consonant deletion, V/W confusion, vowel substitution, R deletion, cluster reduction and voicing errors.

Attempts that reach the phoneme service are scored using the rules in Table 3.3. The service also returns the phoneme accuracy (0-1) and the main error class. An attempt with a known transfer error therefore gets partial credit and is shown to the teacher, instead of being treated as fully correct or as a failure. Because exact and key-phrase matches bypass the phoneme service, those attempts have no error class, and key-phrase matches have no phoneme accuracy value; the trajectory features and the Tier 1 formula treat these as missing rather than inventing a value.

Table 3.3: RC1 scoring rules

Condition	Score
Exact phoneme match, or phoneme accuracy of at least 0.85 with only vowel or R errors	3
Every error belongs to the taxonomy	2 (flagged for the teacher)

Condition	Score
Phoneme accuracy of at least 0.4	1
Otherwise	0

If the service does not respond within two seconds, the system falls back to the earlier fuzzy matching, so a session is never interrupted. Because this fallback silently replaced RC1 throughout the early pilot, a checklist confirming that the service is running is now completed before every session.

3.3.3 RC3: Echolalia detection through response latency

Immediate echolalia is the repetition of speech just heard. It occurs in up to 85% of autistic individuals who develop speech and is usually defined as a repetition within two conversational turns of the original [13]. In Phase 2 the avatar says the word and the child is then asked to say it. A child who echoes the avatar can produce a well-pronounced response without retrieving the word, and could reach "mastered" without having learned it. Published definitions describe immediate echolalia as following the model after only a brief delay [18], but no study reviewed gives a timing threshold, and studies define echolalia inconsistently [19]. RC3 therefore operationalises "immediately after the model" as a measurable response latency and tests the design hypothesis that echoed responses start sooner than genuine productions.

Detection. The app records when the avatar's audio ends and when recording starts. The speech service gives the offset of the first word in the recording, so speech onset is the recording start plus that offset. Response latency is the onset minus the end of the avatar audio. An attempt with a latency under 1,500 ms is flagged as probable echolalia. The threshold is a tunable default and was not recalibrated after the pilot.

Response. When a session pass is echolalic, the word stays in progress: that pass does not count toward Rule 1, but the failure streak is reset so the child is not penalised. Each child's echolalia rate is updated after every attempt as an exponential moving average ($\alpha = 0.3$). When the rate rises above 0.5, the microphone opens three seconds after the avatar finishes, giving the child less chance to echo. Teachers see how often echolalia occurs for each word, because a falling rate is a meaningful sign of progress.

3.3.4 RC2: Learning-trajectory prediction

RC2 predicts whether a child is progressing at a fast, typical or struggling pace on a word, using the features from the child's first session with that word. The prediction adjusts the difficulty ladder (fast lets the child advance after any one easier word is mastered, typical keeps the standard rule, and struggling requires all easier words to be mastered) and the familiarisation dwell for Abilities (for a child predicted to be struggling, the Next button waits for one full animation cycle). Figure 3.3 shows the complete pipeline.

Features. Thirteen features are drawn from the first session:

- the Phase 1 exposure ratio;
- the Phase 2 speech score, phoneme accuracy, phoneme error class, response latency and echolalia flag;
- the Phase 3 response latency, whether the first tap was correct, the number of selection changes (capped at 2 or more) and the number of prompts;
- word difficulty and category;
- a flag for whether Phase 1 applies.

The flag is needed because Abilities words have no counted Phase 1 exposures. For those rows the exposure ratio is set to -1.0, which keeps "not applicable" separate from genuinely missing values. Categorical features are one-hot encoded and re-indexed to a fixed column list stored with the model, so live requests always match the training columns.

Two tier design. Tier 1 is a deterministic formula computed from the same features, with the terms defined in Table 3.4:

$$BaselineScore = 0.35s + 0.25p + 0.15e + 0.15pr + 0.10l$$

Table 3.4: Terms of the Tier 1 baseline formula

Term	Meaning
s	Normalised speech score (a non-verbal score is already binary and is not rescaled)
p	Phoneme accuracy
e	1 minus the echolalia flag
pr	Independence from prompting
l	Fluency based on response latency

The choice of terms and their direction are grounded in the literature: speech correctness [11], echolalia as a step towards learning rather than evidence of independent mastery [19], and reduced prompting as a marker of independent responding [27]. The numeric weights and the class thresholds (fast at 0.75 or more, struggling at 0.40 or less) are informed starting values, not derived ones. When a term is unavailable, its weight is redistributed across the remaining terms rather than filled with an invented value.

Tier 2 is a Random Forest classifier [45]. Its prediction is used only when its confidence, the share of trees voting for the winning class, is above a threshold; otherwise Tier 1's classification is used. Non-verbal attempts always go to Tier 1. The machine-learning path is switched off by default in the product. It was switched on for the field study, where every session is supervised by a teacher and the researcher (Section 3.3.7).

Development history. Tier 2 went through three iterations. They are reported because together they justify the final method.

1. **Synthetic pretraining.** Five algorithms were compared on the 400-row synthetic dataset (100 simulated children) using leave-one-child-out cross-validation with balanced class weights: Random Forest, XGBoost, a support vector machine with a radial basis function (RBF) kernel, Logistic Regression and a multilayer perceptron. Random Forest had the highest macro-F1 and was selected. The deployed model was then trained on the 2,000 pooled rows of the five domain-randomised variants (300 trees, balanced class weights).
2. **Calibration and audit.** The real pilot data was used to calibrate the model's probability outputs instead of retraining it. Oversampling the real data with the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC) was considered and rejected, because it would have created artificial "real" observations. An audit later found that the calibration had been fitted on only nine real rows, one of them struggling. The resulting curve was inverted and made the model predict "fast" for every input, so the calibration step was removed.

3. **Evaluation on real data.** On the 27-row pilot set, the synthetic-trained model did not recognise the struggling class and was below Tier 1 on the verbal-pathway rows (Section 6.1.3). Those pilot struggling labels were later found to be faulty (Section 6.1.4), so the result was not conclusive, but it showed that a model's accuracy on synthetic data says little about its performance on real children. The final Tier 2 model is therefore trained only on real data, and the synthetic-trained model is kept as a comparison (Section 6.1.5).

Final training procedure. The following settings were fixed before the field study data was seen. The treatment of non-verbal attempts was changed when the data was extracted, as explained below and discussed in Section 7.5.

- **Data:** the real rows described in Section 3.2.4; no synthetic rows are used.
- **Classes:** the model has three classes (fast, typical, struggling) only if struggling examples come from at least three different children. Otherwise it has two classes (fast and typical), and struggling detection stays with Rule 2 and Tier 1. This minimum was decided by the researcher, after conducting field sessions with a range of children, and before the real-data model was evaluated. In the final dataset, struggling rows came from 17 accounts (8 of them on the verbal pathway), so the model has three classes.
- **Class imbalance:** balanced class weights, with no oversampling.
- **Model settings:** a small, constrained Random Forest whose settings are fixed in advance rather than tuned, because tuning on so few rows would fit noise: 300 trees, a maximum depth of 4, at least 2 samples per leaf and balanced class weights, with random seed 42. Missing numeric values are filled with medians calculated from the training children only, never from the held-out child; a version that drops incomplete rows instead is also reported. No probability calibration is applied.
- **Non-verbal attempts:** the plan was to exclude them from training. They were kept, because 17 of the 26 struggling rows are non-verbal and excluding them would have left too few struggling examples.

- **Evaluation:** leave-one-child-out cross-validation (Section 3.5), with a majority-class baseline, a label-permutation test and child-level bootstrap confidence intervals.
- **Deployment:** the model is retrained on all real rows and saved with its column list, label encoder, imputation values and the scikit-learn version used. It is intended to replace the synthetic-trained model, but it had not been deployed when this draft was written: the field study continues with the synthetic-trained model (Section 3.3.7). The confidence threshold, provisionally 0.60, is re-set from the leave-one-child-out predictions of the new model: it is the lowest vote share at which Tier 2, on the cases it would handle, reaches a balanced accuracy at least equal to that of Tier 1 on the same cases. If no threshold meets this condition, the machine-learning path stays off and Tier 1 is used. The result is reported in Section 6.1.5.

The Tier 1 fallback and the confidence gate are kept. In the product, the machine-learning-driven difficulty ladder stays off by default until the retrained model has itself been observed in supervised sessions.

Explanations for teachers. Tier 1 predictions are explained by the exact contribution of each formula term. Tier 2 predictions are explained with SHapley Additive exPlanations (SHAP) values [39], summed from the encoded columns back to the 13 named features. Every Tier 2 explanation carries a visible note that it is an early hint from a model trained on little real data, and confidence is shown as a vote share, not a probability. A failed explanation request never changes the prediction shown, and a sweep of 6,482 test inputs confirmed that adding explanations changed no labels.

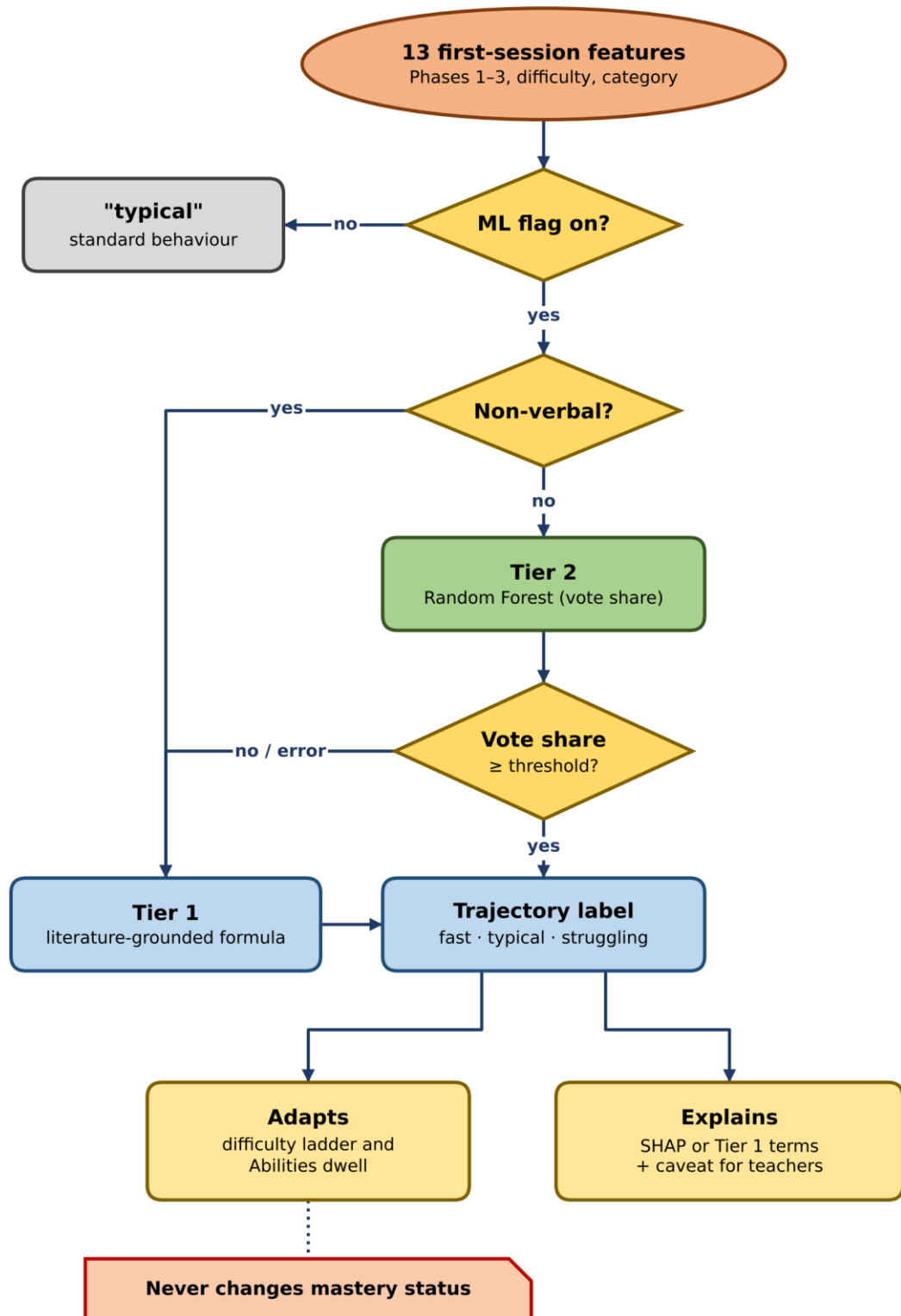


Figure 3.3: RC2 trajectory prediction pipeline

3.3.5 Level 2 sentence construction

Each topic starts with a profile set up by the teacher and child together, recording name, age, home town and favourite activity, plus the friend's or pet's details for those topics. Self-Introduction also includes a "Draw yourself" activity, and the drawing then appears beside the child's sentences. For each sentence, the child:

1. listens to the sentence as it builds up word by word;
2. arranges word tiles into the correct order;
3. fills a single gap;
4. says the sentence.

For Self-Introduction, the child then says all the sentences together as a paragraph before saying them one at a time; Describing a Friend and Describing a Pet use sentence-by-sentence production only. Children who do not speak complete the topic by matching word cards. Spoken sentences go through the same speech assessment as Level 1, with sentence-specific trigger phrases at three levels, from the full sentence down to its key word. An element is counted as produced when its sentence scores 1 or more, matching the school's practice of crediting partial answers. A topic is mastered when the element score is 4 or more in two sessions on different days, and it is marked struggling when the score is 1 or less in three consecutive sessions.

3.3.6 Software verification

Every development task was verified against its acceptance criteria before it was approved. The methods were syntax and module-loading checks for backend code; automated unit tests for the main logic, including the Tier 1 scorer, the trajectory service and the reporting services; and targeted live tests on the device. Database changes were written as reviewed migration scripts and were run manually, never automatically. Chapter 5 gives the details.

3.3.7 Field study procedure

User testing with children began at the partner school in June 2026 and continued with the deployed testing build from 7 October 2026, when the other two schools joined; it is planned to end in the third week of October 2026. Each child at the partner school has completed at least four sessions, and each child at the other two

schools three sessions so far. From 7 October, testing uses the deployed release of the complete AURIVA application, in which all four components, including this one, are integrated into a single app with a shared login and shared student profiles. Children and teachers therefore use the module as it would be delivered, reached through the shared AURIVA interface rather than as a stand-alone prototype. The testing build (auriva-db5c387) was produced from the latest commit of every repository; the only change made during testing was the deployment of the single-voice Level 2 audio (Section 3.2.2). During integration, the component's learning logic, assessment and models were not changed. The only changes were in the frontend: the user interface was restyled to be uniform across the four components, and the Level 2 teacher report was extended to include results from the sentence-by-sentence production, drag-and-drop and sentence-construction activities. The app is installed on two tablets, an Amazon Fire HD 10 (11th generation) and a Samsung tablet, and the backend and the Python microservice run on Azure. Before testing began, the deployment was confirmed to be complete: the backend and microservice are running, speech-to-text and text-to-speech are working, the deployed trajectory model is the uncalibrated Random Forest described in Section 3.3.4, and all database migrations have been applied. Before each testing day, the device is confirmed to be running the testing build, the backend and phoneme service are confirmed to be reachable, and a test session is checked in the database to confirm that scores, latencies and phase records are being stored. These checks were added because of the data-quality problems found in the pilot.

The machine-learning trajectory path is switched on for the field study. The model running during testing is therefore the earlier, synthetic-trained Random Forest (Section 3.3.4), not the final real-trained model: it influences which word each child is offered next and the familiarisation dwell for Abilities, and teachers can view its predictions and explanations in the reports. Because every session is supervised, a teacher can redirect a child at any time, and the prediction never affects mastery status. Labels for retraining come from the mastery rules, not from this model, so its use during testing does not contaminate the training data. Teacher judgements of the

trajectory labels collected during the field study therefore evaluate the deployed synthetic-trained model, and this is stated when they are reported in Chapter 6.

Each child works through Level 1 and, where appropriate, Level 2, with a teacher beside them and the researcher observing. For each attempt, the researcher records on a structured observer sheet:

- the clock time and the word;
- the response type (spontaneous, prompted, echoed, tapped, incorrect or none);
- a pronunciation rating on a 0-2 scale, with any sound error heard;
- whether echolalia was observed;
- the child's engagement on a 0-2 scale.

The pronunciation and echolalia columns are completed before looking at the app's feedback where possible, and a second rater completes an independent sheet when available. The recorded time links each observation to the system's log entry for the same attempt. A session summary records distress, interventions, technical problems and the observer's own view of the child's learning pace.

Each teacher also completes two instruments, both given in Appendix A. These are administered in the remaining field sessions and had not yet been conducted when this draft was written:

- **A task-based usability survey.** The teacher carries out 14 tasks covering profile setup, guiding a child through Levels 1 and 2, and checking progress. They rate each task's ease on the Single Ease Question (SEQ, 1-7) [46] while the researcher records completion. They then rate nine statements about supporting the child, complete the System Usability Scale (SUS) and answer open questions.
- **A feedback questionnaire.** It covers suitability for the children, agreement with the app's speech and echolalia judgements, and the trajectory reports. Before seeing the app's label for each child, the teacher writes down their own judgement of that child's learning pace.

Field study data is used in two ways: to evaluate RC1, RC3 and usability, and, once labelled by the mastery rules, as training and evaluation data for the final trajectory model.

3.4 Measures and Evaluation Metrics

Table 3.5 summarises the measures used for each part of the component.

Table 3.5: Evaluation measures by component

Component	Measure	Source
Software correctness	Pass or fail against each task's acceptance criteria; automated tests	Task specifications, test runs
RC1 speech assessment	Agreement between the system's score and the human pronunciation rating; agreement on the error type	System log matched to observer sheet
RC3 echolalia detection	Precision and recall of the echolalia flag against observed echoing; distribution of response latencies	System log matched to observer sheet
RC2 trajectory prediction	Accuracy, balanced accuracy, macro-F1 and per-class recall (struggling recall as the main clinical measure, if the model has three classes), for Tier 1, the synthetic-trained model and the real-trained model on identical folds	Leave-one-child-out cross-validation
RC2 in the field	Agreement between the system's label and the teacher's own judgement	Teacher questionnaire
RC2 explanations	Teacher ratings of how clear and useful the explanations were	Teacher questionnaire
Teacher usability	Task completion (unaided, with help, not completed); median ease per task; SUS score (0-100)	Task usability survey
Child suitability	Engagement, points where engagement dropped, distress events, session completion; teacher suitability ratings	Observer sheet, teacher questionnaire
Reliability and integration	Sessions without technical failure in the deployed, integrated app; problems when moving between AURIVA components; longest wait for feedback	Observer sheet, system logs

Two mappings are fixed before any field data is analysed:

- **RC1 scale mapping.** The human scale (0-2) and the system scale (0-3) differ, so both are reduced to the same practical decision: accepted, or needs more practice.
- **RC3 ground truth.** Phase 2 asks the child to repeat the avatar's word, so the observer sheet does not count a correct repetition of the model as echolalia. Following a published operational definition in which repeating the prompt itself ("say boat") counts as echolalia while the target alone ("boat") does not [19], an attempt is counted as an observed echo when the child repeats the prompt

wording or the avatar's carrier phrase, or copies the avatar's intonation within the same turn; producing the target word alone is not.

For RC2, balanced accuracy and macro-F1 are reported alongside raw accuracy, because on a small, unbalanced set raw accuracy can look high even when a model is at chance level for the minority class. The proposal's numeric targets (at least 75% speech assessment accuracy, at least 70% task completion, and a teacher satisfaction score of at least 4.0 out of 5) are reported in Chapter 6 as reference points rather than pass or fail criteria, because the design and the sample sizes changed.

3.5 Data Analysis

Field study. The quantitative results are analysed descriptively, using counts, percentages, medians and ranges, with the sample size stated beside every figure. No significance tests are applied, because they would not be meaningful at this sample size.

RC1 and RC3. Each system log entry is matched to the observer's record of the same attempt by time. Agreement is reported as a percentage, and also as Cohen's kappa if enough attempts are collected.

RC2. The model is evaluated with leave-one-child-out cross-validation on the real dataset. In each fold, the model is trained on every child except one and tested on that child's rows, so no child appears in both training and testing. Predictions from all folds are pooled to compute the metrics, and Tier 1 and the earlier synthetic-trained model are scored on the same rows. Results are reported for all rows and for the verbal rows alone, because non-verbal attempts always go to Tier 1 in the product. Every prediction is made one row at a time, exactly as in the live system; an earlier evaluation script encoded many rows together, which corrupted one feature column, and this was found and fixed (Chapter 5). Performance on the training rows is never reported. Results that depend on data collected after 11 October 2026 are marked as pending in this draft.

Usability. SUS scores are calculated with the standard method and reported per teacher, with the mean and range, against the commonly cited average of 68. For each task, completion counts and median ease are reported.

Open-ended answers. Answers to the open questions and the observer notes are grouped into themes, and representative quotes are used to explain the quantitative results.

3.6 Ethics, Privacy, Security and Responsible AI

Consent and participation. The Ministry of Education of Sri Lanka issued a written letter authorising the research group to visit schools in the Colombo district and collect the data needed, and each participating school gave its permission after receiving this letter. Each child's willingness to take part is respected throughout every session. A teacher is present at every session and can stop it at any time, and a session ends immediately if a child shows distress. Teachers take part voluntarily and anonymously, and consent is confirmed for every child before each testing day.

Child-safe design. The module avoids features that pressure children:

- no failure messages or scores shown to the child;
- no time limits;
- feedback that is always gentle and graded;
- a non-verbal pathway, so children who do not speak can take part and progress fully;
- a cap on retry loops and adaptive wait times, to prevent repetitive or aversive cycles.

Privacy and data handling.

- Children are identified only by codes in all research data, and no video of children is recorded.
- Personal details used for Level 2 personalisation are stored separately from interaction data and are used only to generate the child's own audio.
- Spoken responses are sent to Google Cloud Speech-to-Text for transcription through its synchronous recognition service. Data logging is disabled for the project, so Google processes the audio in memory and does not store it or use it to improve its models. The backend does not save the audio either; only the text transcript is kept in the database.

- The training data contains the internal database identifier of each child and no other identifying information. This was confirmed to be covered by the existing ethics approval.
- Interaction data is stored in the AURIVA group's shared PostgreSQL database, which is reached over an encrypted connection and is accessible only to the four members of the AURIVA research group. Only anonymised, aggregated results are reported.
- The Google Cloud project used for speech-to-text is owned by the researcher, and its service-account key was added to the integrated deployment manually rather than through the shared repositories, and it has not been committed to version control.

Responsible AI. The trajectory model was designed so that it cannot harm a child's recorded progress:

- it cannot change mastery status;
- when it is unsure, the literature-grounded formula is used;
- its influence on the difficulty ladder is switched off by default, and was switched on only for the supervised field study;
- teachers see an explanation for every prediction, with a clear statement of its limits.

Reported model performance always comes from children the model was not trained on. The synthetic data was never presented as real, and it is not part of the final model. The test-account rows in the real dataset are disclosed (Section 3.2.4). Weaknesses found during development, including the corrupted struggling labels and the faulty calibration, are reported openly. AI-generated teaching media was reviewed before use, and AI assistance in development is declared in Appendix B.

3.7 Reproducibility and Research Artefacts

The component was developed in three code repositories: the mobile app (React Native with Expo), the backend (Node.js and Express with PostgreSQL), and the Python microservice for phoneme scoring, trajectory prediction and explanations. For integration, the app and backend code were merged into the AURIVA group's

shared app and backend repositories, where each member worked on a personal branch (this component's backend branch is dev/maryse), and all four components share one database. Each member's Python microservice remains a separate repository; this component's is auriva-microservice. The testing build auriva-db5c387 was produced from the latest commit of every repository. The planning records are kept beside the repositories: the status file, the numbered decision log, the roadmap, every task specification with its acceptance criteria, and the audit reports. Any design decision in this thesis can therefore be traced to the record that made it. For the machine learning work:

- The training notebook is kept as evidence, including the synthetic iteration.
- The final model is produced by one training procedure run on the real dataset.
- The model file stores the trained model with its column list, label encoder and scikit-learn version.
- Library versions are pinned (scikit-learn 1.6.1, joblib 1.4.2) and the random seed is fixed at 42. The synthetic generator used seed 42 and its variants seeds 42 to 46.
- The faulty calibrated model is kept under a separate name as evidence for the audit.
- Runtime settings (the machine-learning flag and the confidence threshold) are recorded in this thesis, because the environment configuration file is not stored in the repositories.
- The training notebook, with its saved outputs, records the algorithm comparison, the pretraining and the calibration step reported in Chapter 6. A separate five-fold comparison of seven algorithms, run during development, is not in the repositories and is therefore not reported as a result.

The observer sheet and teacher instruments are given in Appendix A. The interaction data cannot be shared: it is held only in the AURIVA database, accessible to the research group for implementation and testing.

3.8 Commercialization Aspect of the Product

AURIVA is commercialized as one product, the Adaptive English Language Learning Platform for Autistic Children in Sri Lanka, of which this component is one part. This section summarizes the commercial model presented by the group and explains this component's role in it.

3.8.1 Value proposition and target users

The platform has two customer personas, summarized in Table 3.6. The paying customer is the institution, a special education school or institute, represented by its principal or special education coordinator; parents benefit indirectly as their children's communication and independence improve.

Table 3.6: Customer personas

Persona	Challenges	Needs and goals
Autistic child: learner with ASD, estimated age 5-15	Limited verbal and non-verbal communication; English as a second language; difficulty with abstract or unsupported instruction; fine motor difficulties with handwriting	Structured, visual and repetitive content; clear feedback and consistent pacing; a non-verbal participation pathway when needed
Special educator: special education teacher, estimated age 21-60	No data driven tools to track individual progress; specific learning gaps are hard to identify manually; needs curriculum-aligned, ready to use content	Monitor each child's scores and trends at a glance; identify which children need additional support; deliver structured lessons without technical expertise

This component answers several of these needs directly. For the child, it provides structured, visual and repetitive teaching of everyday words and sentences, consistent pacing adjusted by the trajectory prediction, gentle graded feedback, and a non-verbal pathway through every activity. For the teacher, it provides curriculum-traced content that is ready to use, speech assessment that expects the pronunciation patterns of Sinhala speakers and does not mistake echoing for learning, and reports that show each child's progress and explain which children are struggling and why.

3.8.2 Market and competitive position

The initial market is Sri Lanka's special education sector, which the group identified as 27 special education schools and 704 special education institutes. Existing products (Table 2.1) are either communication aids, such as Proloquo2Go and Avaz, or general speech and language applications, such as Speech Blubs and Mita. None

follows the Sri Lankan curriculum, assesses Sinhala-influenced English, or explains its decisions to teachers, and local research platforms target general skills rather than spoken English dialogue [37]. The platform's advantage is its local fit: curriculum-traced content, culturally familiar scenarios, Sinhala instructions that can be switched off for each child, and names and home towns pronounced correctly in the child's own sentences. Because the four components are sold together, a school gets concept learning, handwriting, pronunciation and dialogue in one application with one login and one shared student profile. The main competitive risks are larger international applications adding local languages and the limited budgets of government schools.

3.8.3 Business and sustainability model

The platform is offered as software as a service, licensed per institution rather than per student, so that costs are predictable for schools working within fixed allocations. Table 3.7 shows the revenue model.

Table 3.7: Revenue model

Offer	Capacity	Target segment
Freemium trial	Free trial	Any institution wishing to evaluate the platform before committing
Tier I	Up to 30 students	Small special education units
Tier II	30-120 students	Mid-scale schools and larger special education units

The free trial suits the way Sri Lankan public schools usually need evidence from a trial before a budget is approved. For rural and low-resource schools, reduced or grant-supported pricing can be sought with non-governmental organisations and government inclusive-education programmes. Running costs for this component are cloud hosting for the backend, database and microservice (currently Azure), speech-to-text fees for each minute of child speech, and text-to-speech fees for personalised sentences; the latter can be reduced by caching each child's sentences, which change only when the profile changes. Recurring costs include maintenance, security monitoring and backups, teacher onboarding, and content updates when the curriculum changes.

The product is designed to improve as it is used. With consent, interaction data from subscribing schools can be used to retrain the trajectory model, which can be

retrained on real data alone as it grows (Section 3.3.4). Word sets, scenarios and voice settings are stored as data, so they can be updated without code changes, and short teacher-led onboarding keeps the cost of introducing the platform low.

3.8.4 Scalability, intellectual property and compliance

The backend keeps no session state in memory, so more instances can be added as the number of schools grows; each microservice can be scaled independently; and the database can be scaled vertically or with read replicas for reporting. The main limits are the quotas and costs of the external speech services. The core intellectual property of this component is its learning-flow and mastery logic, the Sinhala error taxonomy and scoring rules, the echolalia rule, the trajectory models trained on real data, and the curriculum-traced content set. It is delivered only through the hosted service, not by distributing source code, and third-party services (Google Speech-to-Text and Azure Text-to-Speech) are used under their own licences and kept separate from the product's own logic. Because the product handles children's data, children are identified only by codes in research data, personal details are stored apart from interaction data, access is limited to authorised staff, and speech data logging is disabled with the speech provider. A commercial release would also need compliance with Sri Lanka's Personal Data Protection Act No. 9 of 2022, data-processing agreements with each institution, and recorded parental consent for any data used to retrain the model.

3.8.5 Contribution to the Sustainable Development Goals

The product supports Sustainable Development Goal 4, Quality Education, particularly Target 4.5 on equal access to education for persons with disabilities and Target 4.a on inclusive learning environments, through curriculum-aligned teaching, a non-verbal pathway and a design with no failure state for the child. It also supports Goal 10, Reduced Inequalities, particularly Target 10.2 on the social inclusion of all irrespective of disability, by helping autistic children gain the everyday English on which education and later employment in Sri Lanka depend.

4 PRODUCT AND SOLUTION DESIGN

This chapter describes the design of the Adaptive Dialogue Learning and Progression Tracker as it was built and deployed. It states the requirements the design had to meet and how far each was met (Section 4.1), places the component within the integrated AURIVA application (Section 4.2), records the main design alternatives and the reasons for each choice (Section 4.3), and then describes the detailed design of every part of the component, including the three research contributions (Section 4.4). Section 4.5 summarises the standards and professional practices followed. Each design choice is linked back to the research problem: autistic Sri Lankan children need to learn functional English dialogue words in a way that tolerates their speech patterns, adapts to their pace and gives teachers evidence they can trust.

4.1 Requirements

4.1.1 Stakeholders

The proposal identified three stakeholders. The **autistic child** is the primary end user, who follows the learning content and attempts the activities. The **special education teacher** acts as a proxy user: the teacher sits with the child, operates the application on the child's behalf where needed, and uses the reports. The **principal** acts as the system administrator. Classroom observation during the school visit showed that sessions are always teacher-assisted, so the child-facing screens address the child while small helper text addresses the teacher.

4.1.2 Requirements from classroom observation

Besides the requirements written in the proposal, several requirements came directly from the partner school: from the teachers' guidance and from classroom observation during the visit on 2 April 2026 and later sessions. These are listed in Table 4.1 with the design feature each one produced.

Table 4.1: Requirements from the partner school and classroom observation

Observation or guidance	Design response
Some children did not speak at all, or stopped speaking after a refusal	A non-verbal pathway through every phase, so a child can progress and master words by tapping pictures

Observation or guidance	Design response
Children who refused to speak were kept waiting at the microphone	Adaptive wait times: consecutive refusals in a day shorten the waits (1.0, 0.7, 0.5), and after three refusals the child goes straight to the non-verbal pathway for that day
Children tapped the record button repeatedly and quickly	A guarded recorder that ignores rapid double-taps, so overlapping recordings cannot start
Children needed more time to recognise a word before saying it	Two extra familiarisation screens (animated word, then bold static word) before Phase 2
Children became upset by being marked wrong	Assessments with no failure state: gentle, graded feedback and celebration regardless of score; the score is kept for the teacher
Children engaged most with interactive, hands-on activities	Interactive activities preferred in new designs, such as drag-to-line, word tiles and the "Draw yourself" canvas
Teachers credit attempts and partial answers, not only complete ones	Partial credit in the speech score (0-3) and element-based scoring in Level 2
Teachers sit with the child during every session	Teacher-assisted design: profile setup done together, and an adult check before backward navigation across a phase boundary

4.1.3 Functional requirements

The proposal listed 19 functional requirements (FR). The design changed during development, most importantly when Level 3 (Conversation Builder) was removed after test rounds with children showed that conversation was beyond their current level (Section 3.1). Table 4.2 gives the final status of each requirement. "Shared AURIVA platform" means the requirement is met by the integrated application that all four components share, not by this component's own code.

Table 4.2: Functional requirements and their final status

FR	Requirement (summarised from the proposal)	Status	How it is met or why it changed
FR1	The administrator registers and manages child and teacher profiles	Met	Shared AURIVA platform: one login and shared student profiles for all four components
FR2	The administrator assigns students to a teacher	Met	Shared AURIVA platform
FR3	Content uses only words from the Sri Lankan English curriculum (Grades 1-3)	Met	Level 1 words and Level 2 activities trace to the Grade 1-2 activity book; Level 2 topics match the school's Grade 3 speaking criteria (Table 4.7)

FR	Requirement (summarised from the proposal)	Status	How it is met or why it changed
FR4	Level 1 follows the three-phase method: Familiarisation, Production, Contextual Awareness	Met	Section 4.4.1
FR5	The LLM generates conversation turns using only mastered words	Withdrawn	Level 3 was removed; the effort moved to Levels 1 and 2 and to the trajectory model (RC2)
FR6	Clear interfaces with large touch targets	Met	Large picture targets and simple screens; the user interface was made uniform across the four components during integration
FR7	The child speaks by tapping the microphone, and the audio is captured	Met	Guarded recorder; audio is sent to the backend for assessment (Section 4.4.2)
FR8	Child-friendly feedback with avatar animations and simple phrases	Met	Graded feedback with no failure state
FR9	Difficulty adapts to the child's previous performance	Met	Difficulty ladder driven by mastery and by the RC2 trajectory prediction (Sections 4.4.3 and 4.4.6)
FR10	A learning path for non-verbal children after several prompts	Met	Non-verbal pathway after three score-0 attempts, plus adaptive wait times
FR11	The teacher logs in and views assigned students	Met	Shared AURIVA platform
FR12	The teacher views each student's progress and mastered words across all modules	Met	This component's word and topic progress appears in its teacher reports; the cross-module view is provided by the shared platform
FR13	A detailed, dated view of every attempt and result	Met	Attempt history in the teacher reports (Section 4.4.7)
FR14	A summary of progress at each level with visual indicators	Met	Status indicators and trend graphs
FR15	The teacher adjusts session frequency and daily usage for each student	Partly met (changed)	There are no per-student frequency or daily-usage settings. Instead, the teacher can pause or exit a session at any point where needed, which suits the supervised classroom setting
FR16	Printable, explained reports aligned with the syllabus	Met	Printable trajectory and Level 2 reports with plain-language explanations
FR17	The principal views high-level reports of each student	Met	Shared AURIVA platform

FR	Requirement (summarised from the proposal)	Status	How it is met or why it changed
FR18	The principal monitors system usage by teachers and students	Met	Shared AURIVA platform
FR19	A contrastive Sinhala SOV / English SVO instruction phase before sentence production	Met (moved)	Moved from Level 3 to Level 2, where sentence production now happens (Section 4.4.8)

SOV and SVO stand for Subject Object Verb and Subject Verb Object word order. Several features were added that the proposal did not list: the three research contributions (phoneme-aware speech assessment, learning-trajectory prediction with explanations, and echolalia detection), formative evaluations, the periodic production probe for non-verbally mastered words, combined Sinhala and English instruction audio, and the "Draw yourself" activity.

4.1.4 Non-functional requirements

Table 4.3 lists the non-functional requirements (NFR) from the proposal and how the design addresses each one.

Table 4.3: Non-functional requirements and design response

NFR	Requirement	Design response
Security	Student data in a password-protected database, accessible only to authorised people, with personal data separate from interaction data	Shared PostgreSQL database over an encrypted connection, accessible only to the four group members; Level 2 personal details kept in their own table, apart from attempt data; credentials kept out of version control
Performance	Speech assessment feedback in under 5 seconds	Synchronous speech-to-text for short clips; a 2-second timeout on every microservice call, with a local fallback
Performance	LLM turn generation within 10 seconds	No longer applies: Level 3 was removed
Usability	Simple, high contrast, large visuals for the child; reports a non-technical teacher can understand	Large pictures and minimal text; plain language report explanations; evaluated with the SUS and task survey (Chapter 6)
Reliability	Sessions continue if an external service is unavailable	Every external dependency has a fallback: fuzzy matching if the phoneme service fails, the Tier 1 formula if the trajectory model fails, and "typical" (the unchanged behaviour) if features cannot be built
Scalability	Concurrent sessions for up to 20 children	Stateless backend and microservice requests

NFR	Requirement	Design response
Accessibility	Usable on Android and iOS devices	Built with React Native and Expo, which target both; tested on Android-based tablets only

4.2 Component Architecture within the Complete Solution

4.2.1 Place in the AURIVA application

AURIVA is a four component adaptive English learning platform for autistic children in Sri Lankan special education. The four components were developed separately and then integrated into a single mobile application with one login, shared student profiles and one shared PostgreSQL database. Each member's machine-learning and speech services run in that member's own Python microservice. Figure 4.1 shows the integrated architecture and the boundary of this component.

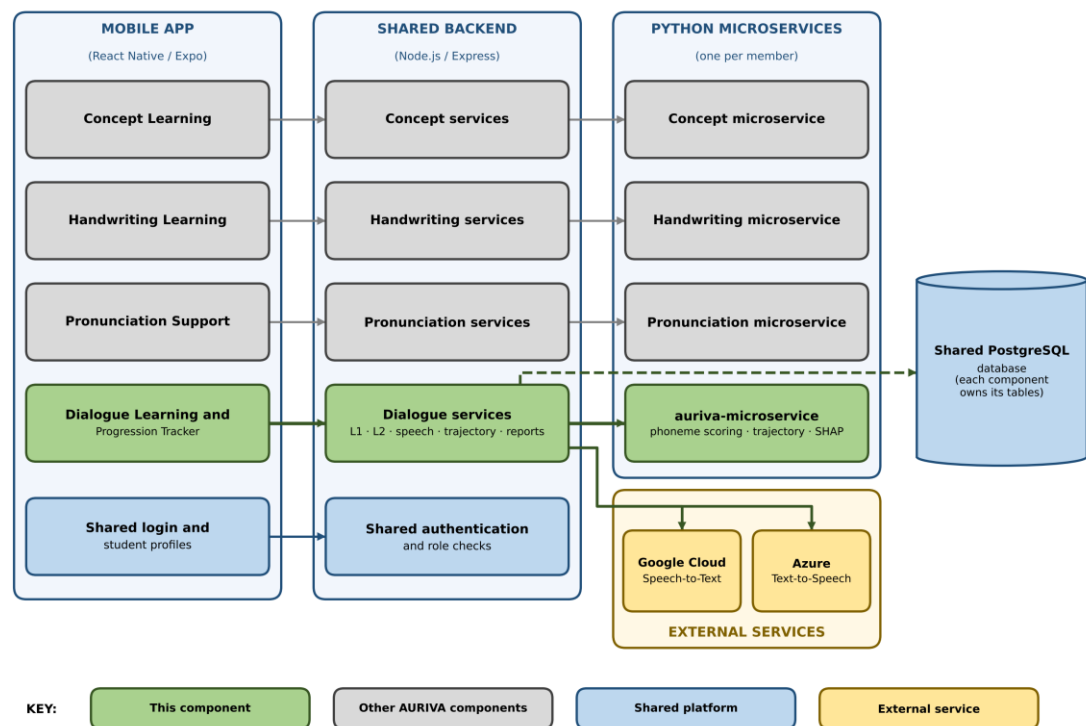


Figure 4.1: Integrated AURIVA architecture and the boundary of this component

This component owns its own screens in the app, its backend services and routes, its database tables (Section 4.4.9) and the auriva-microservice service. It uses three things provided by the platform: authentication, the shared student profile and the shared navigation into each component. During integration only the frontend was

changed: the screens were restyled to match the other components, and the Level 2 teacher report was extended to include sentence-by-sentence production, drag-and-drop and sentence-construction results. The learning logic, assessment and models were not changed.

4.2.2 Internal architecture

Inside its boundary the component has five layers, shown in Figure 4.2:

1. **Presentation layer.** The child-facing screens for Level 1 and Level 2, and the teacher-facing report screens, built in React Native with Expo.
2. **Application programming interface (API) layer.** REST routes in the Node.js and Express backend, protected by the platform's authentication.
3. **Service layer.** The backend services that hold the learning logic: the dialogue service (Level 1 flow and mastery rules), the speech assessment service, the trajectory service with the Tier 1 scorer, the Level 2 service, the evaluation service and the report services.
4. **Intelligence layer.** The Flask microservice (port 5001) with four endpoints: /health, /score-phoneme (RC1), /predict-trajectory (RC2 Tier 2) and /explain-trajectory (SHAP explanations).
5. **Data layer.** The component's tables in the shared database, accessed through the Sequelize object relational mapper, with every schema change made by a reviewed migration.

Two external services are used: Google Cloud Speech-to-Text (STT) for transcription, called from the backend, and Azure Text-to-Speech for personalised Level 2 sentences. Keeping the learning rules in the backend and the models in a separate microservice means the mastery logic never depends on the models: if the microservice is down, the backend still decides every outcome.

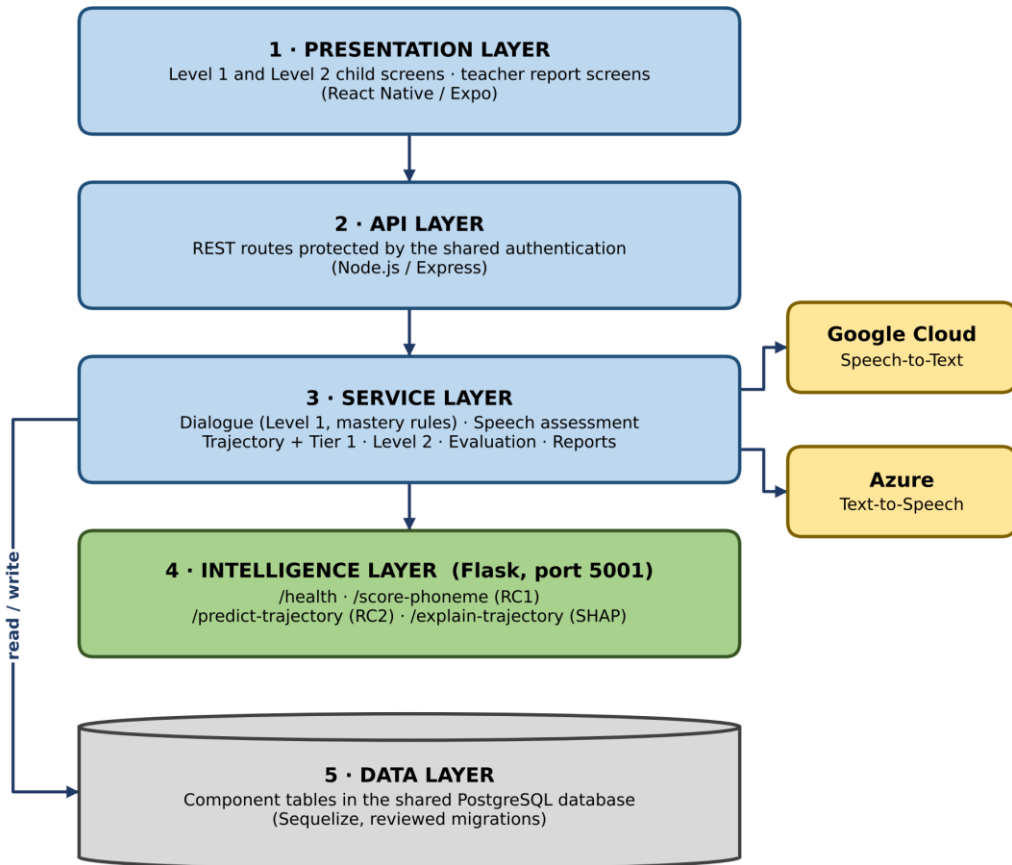


Figure 4.2: Internal layered architecture of the component

4.2.3 Main data flows

Assessing a spoken attempt. Figure 4.3 shows the sequence for one Phase 2 attempt. The app records the child's response and sends it with the time the avatar's audio ended and the time recording started. The backend converts the audio to uncompressed WAV, sends it to the speech service, and receives the transcript with the timing of each word. It first checks the transcript against the word's accepted forms; if there is no match it calls `/score-phoneme`. It then computes the response latency and the echolalia flag, stores the attempt, updates the child's echolalia rate, and returns the score and the next step to the app. If the phoneme service does not answer within two seconds, the fuzzy-matching fallback produces the score instead.

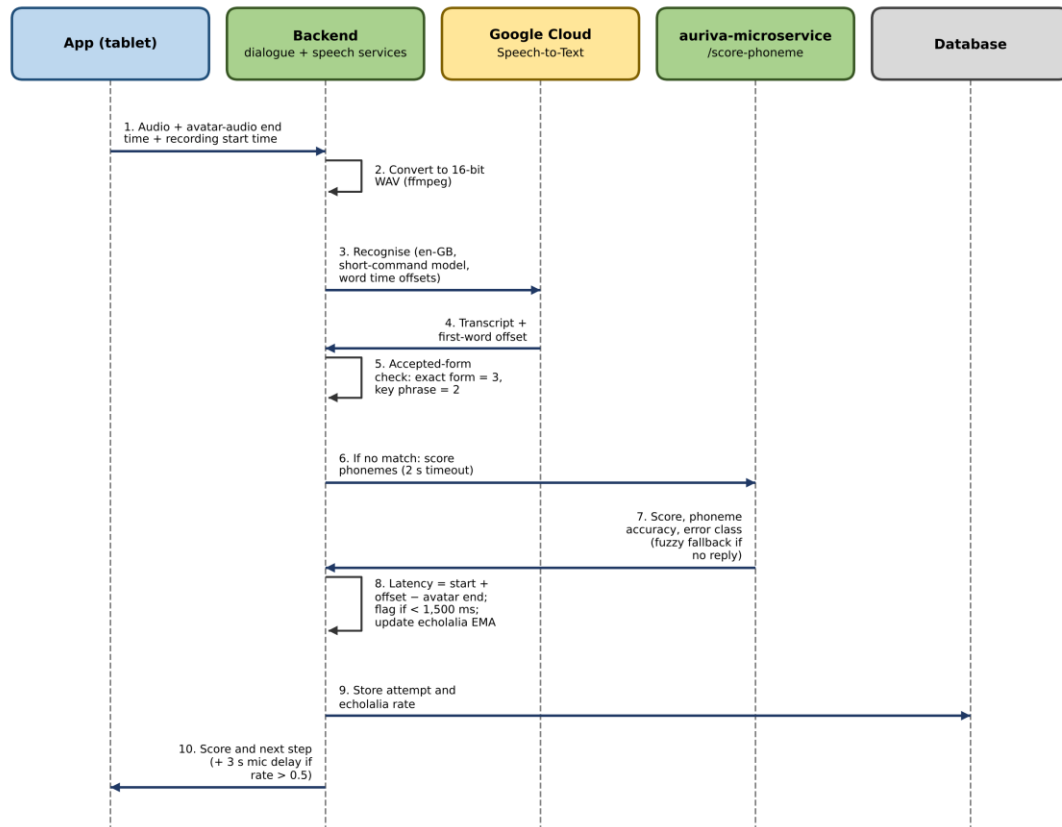


Figure 4.3: Sequence of a Phase 2 spoken attempt

Choosing the next word. When a child opens a category, the backend selects the next word. If the machine-learning flag is on, it assembles the 13 first-session features for each relevant word, asks /predict-trajectory for a prediction, and applies the confidence gate and the Tier 1 fallback (Section 4.4.6). The resulting label sets the difficulty gate for the next word. If the flag is off, or if the features cannot be assembled, the label is "typical", which keeps the standard behaviour.

Building a Level 2 session. When a topic starts, the backend reads the child's profile for that topic, builds the target sentences, produces the audio for each sentence with Azure TTS, and returns the sentences, word tiles, gap options and audio to the app.

4.3 Design Alternatives and Decisions

Every significant design decision was recorded in a numbered decision log with its reasons. Table 4.4 summarises the decisions that most shaped the component, the alternatives considered and the reason for the choice.

Table 4.4: Main design decisions

Decision	Alternatives considered	Reason
Remove Level 3 (LLM Conversation Builder)	Keep Level 3 with a seed-library fallback	Test rounds showed children could learn words but not hold conversations, and unpredictable generated dialogue risked confusing or unsettling them; effort moved to Levels 1 and 2, trajectory prediction and echolalia detection
Remove the Days of the Week category	Keep it as a fourth Level 1 category	Sequential recall of day names does not exercise pragmatic, context-dependent use of a word
Phoneme-aware speech assessment (RC1)	Fuzzy string matching only, as in the proposal	String matching rewards the Sinhala L1 transfer errors the child needs help with (Section 4.4.4)
Accepted-form check before phoneme scoring	Send every transcript to the phoneme scorer	Exact and key-phrase matches need no alignment, which keeps feedback fast
Fuzzy matching kept as a fallback, with a 2-second timeout	Fail the attempt if the service is down	A session must never be interrupted by a service failure
Echolalia detection by response latency (RC3)	No detection; acoustic comparison with the avatar's audio	Latency is available from data already collected, needs no extra model and is explainable to teachers
An echolalic pass gives no mastery credit but resets the fail streak	Count it as a pass; count it as a fail	Avoids crediting an unretrieved word without penalising the child
Mastery decided by fixed rules; the model only adjusts difficulty and dwell	Let the model decide mastery	Labels must come from a stable source, and a model error must not change a child's recorded progress
Two-tier hybrid trajectory prediction	A single model with a bare fallback; the formula only	Tier 1 gives an explainable, literature-grounded answer whenever the model is unsure or unavailable
Random Forest for Tier 2	Gradient Boosting, Logistic Regression, Decision Tree, Naive Bayes, SVM, k-nearest neighbours	Best result in the algorithm comparison; works with mixed feature types and supports tree-based explanations
Machine-learning path behind a flag, off by default	Always on	The product must not depend on the microservice or on an unproven model
Confidence shown as a vote share, never a probability	Show a calibrated probability	The calibration was found to be faulty and was removed, so a probability would overstate certainty
SMOTENC oversampling rejected	Oversample the real data	It would create artificial "real" observations; agreed with the supervisor
Calibration step removed	Keep the calibrated model	An audit found it had been fitted on nine rows and predicted "fast" for every input

Decision	Alternatives considered	Reason
Final model trained only on real data	Keep the synthetic-trained model	Accuracy on synthetic data did not reflect performance on real children, and the product should learn from its own learners; the synthetic-trained model is kept as a comparison
A -1.0 value for the Abilities exposure ratio, with an applicability flag	Treat it as missing	"Not applicable" is different from "missing" and must stay separable for the model
Phase 1 exposure ratio recorded once, at the Phase 1 gate	Use the live counter	The live counter resets after every session, so only the snapshot reflects the child's first-exposure pace
Level 2 activities traced to textbook pages	Activities designed freely	Ecological validity and familiarity for teachers and children
Pet picker limited to the six animals in the textbook	An open list with an "other" option	Matches the activity book exactly and keeps validation and distractors clean
Self-portrait stored as stroke data, not as an image	Capture and upload an image	Simpler, and it can be redrawn at any size anywhere it appears
Human-recorded Sinhala instruction audio, with a per-child switch	Sinhala TTS	Natural voice quality for children, offline reliability, and support that can be faded
Single Level 2 TTS voice with Sinhala names marked inside it	Two voices, one English and one Sinhala chosen by gender	Consistent voice across a sentence; the newer voice could pronounce Sinhala names within the English sentence
Database changes before interface changes; one category at a time	Build screens and data together	Each change could be verified alone, and risky changes were piloted on one category first

Three points connect these decisions. First, every part that uses a model or an external service has a deterministic fallback, so the child's session never depends on it. Second, the model can influence the order and pace of learning but never the record of what the child has learned. Third, wherever data was thin, the design chose an honest, simpler option over a more impressive one: the synthetic data and the calibration were both dropped once they were shown not to hold on real data.

4.4 Detailed Design

4.4.1 Level 1 learning flow

Level 1 teaches the 30 words in Table 3.1. The child chooses a category, and the backend offers the next word allowed by the difficulty ladder. Each word passes through the three phases described in Section 3.3.1. Figure 4.4 shows the main screens for the word "Thank You".

- **Phase 1 (Familiarisation).** For Greetings and Magic Words: scenario videos, then the animated-word and bold-word screens, then drag-to-line, which places the word in a speech bubble. Each exposure is counted automatically against the child's required count. For Abilities: a single tap-to-see interaction and drag-to-line, with no counted exposures. A rewatch cap allows the video to be replayed at most once if the child struggles twice in the same sitting.
- **Phase 2 (Production).** Described in Section 4.4.2.
- **Phase 3 (Contextual Awareness).** Picture scenarios A, B and C, with a checkpoint scenario added when A is wrong. The app records, for each scenario, the response time, whether the first tap was correct, the number of selection changes and the number of prompts. These records feed the trajectory model.

Two further features complete the flow. **Formative evaluations**, in which the child matches words to pictures, unlock once at least three words in a category are mastered and use at most five pairs; they have no fail state and the score is kept only for the teacher. An **adult check** is shown before the app navigates backwards across a phase boundary, so a child cannot accidentally undo progress.

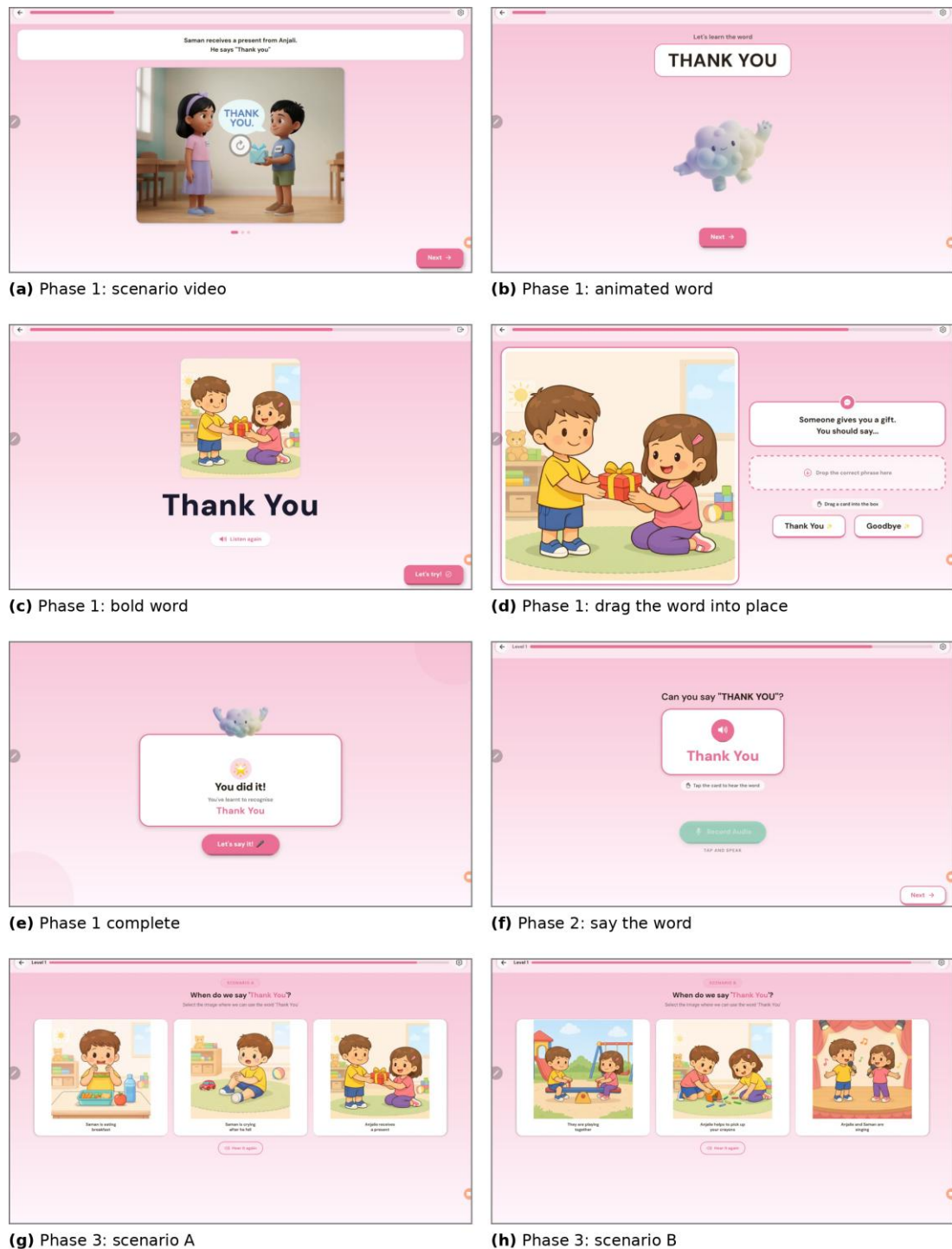


Figure 4.4: Level 1 screens for one word ("Thank You")

4.4.2 Production phase and prompt hierarchy

Phase 2 is where most of the component's design work meets. The avatar models the word, and the child taps the microphone to answer. The record button uses a guarded recorder, so a second tap while a recording is starting or stopping is ignored. The app

records the time the avatar audio ended and the time recording started, which RC3 needs.

If the attempt scores 2 or more, the child moves to Phase 3. Otherwise the prompt hierarchy escalates:

- **Attempt 1:** the avatar models the word at normal speed.
- **Attempt 2:** the difficult letters of the word are highlighted (each word has a fixed cue, for example the "Th" in "Thank you"), and the word is replayed at 60% speed.
- **Attempt 3:** a "Say it with me!" prompt, in which the child speaks together with the avatar.

The prompt sequence is fixed for each word and is not driven by the RC1 error class; using the error class to choose the prompt is left as future work. Three consecutive attempts with no recognisable speech (score 0) switch the child to the non-verbal pathway, where the same goal is met by tapping the correct picture. Refusals across the day shorten the wait before each prompt, and three refusals send the child straight to the non-verbal pathway for the rest of that day; any speech attempt resets this. A session passes when Phase 2 reaches its threshold (2 or more on the verbal pathway, 1 on the non-verbal pathway) and Phase 3 is passed.

4.4.3 Mastery rules and progression

The status of each word for each child is held in one progress record and changes only through the five rules in Section 3.3.1. Figure 4.5 shows the resulting states and transitions. The design keeps two separate counters: a session-level count of consecutive failed sessions, used by Rule 2, and a within-session count of score-0 attempts, used to trigger the non-verbal pathway. An earlier version used one counter for both, which let the failure count exceed the number of sessions; this fault corrupted the pilot struggling labels (Section 3.2.3) and was fixed by separating the counters.

Mastery requires two passed sessions on different calendar days, compared as UTC (Coordinated Universal Time) dates. An echolalic pass leaves the word in progress, does not count toward mastery and resets the failure streak. The progress record also

stores how the word was mastered (verbal, non-verbal or mixed) and the date of the last production probe. Rule 4 adjusts the number of Phase 1 exposures the child needs: it increases by one, up to six, after three or more sessions without a pass. The code also contains a branch that would reduce the count by two after a word is mastered very quickly, but that branch cannot be reached in the current flow, so in practice Rule 4 only increases exposures. This is reported as a limitation in Chapter 7.

Difficulty ladder. Words within a category are ordered by difficulty. The next word is offered from the lowest difficulty that still has unmastered words, unless the gate to the next difficulty is open. The trajectory label changes the gate: for "fast", any one mastered word at the lower difficulty opens it; for "typical", the standard session-unlock rule applies; for "struggling", every word at the lower difficulty must be mastered first. A struggling word returns once the easier words are mastered (Rule 3).

Production probe. Every 14 days, a word mastered through the non-verbal pathway can be offered as a probe through a dismissible banner on the category screen. The child is invited to say the word, with no prompt escalation. If speech emerges, the mastery path becomes "mixed" and the child is celebrated. If not, a simple picture check follows, and any tap ends with the same neutral message. A probe only writes the probe date and the mastery path; it never changes the word's status.

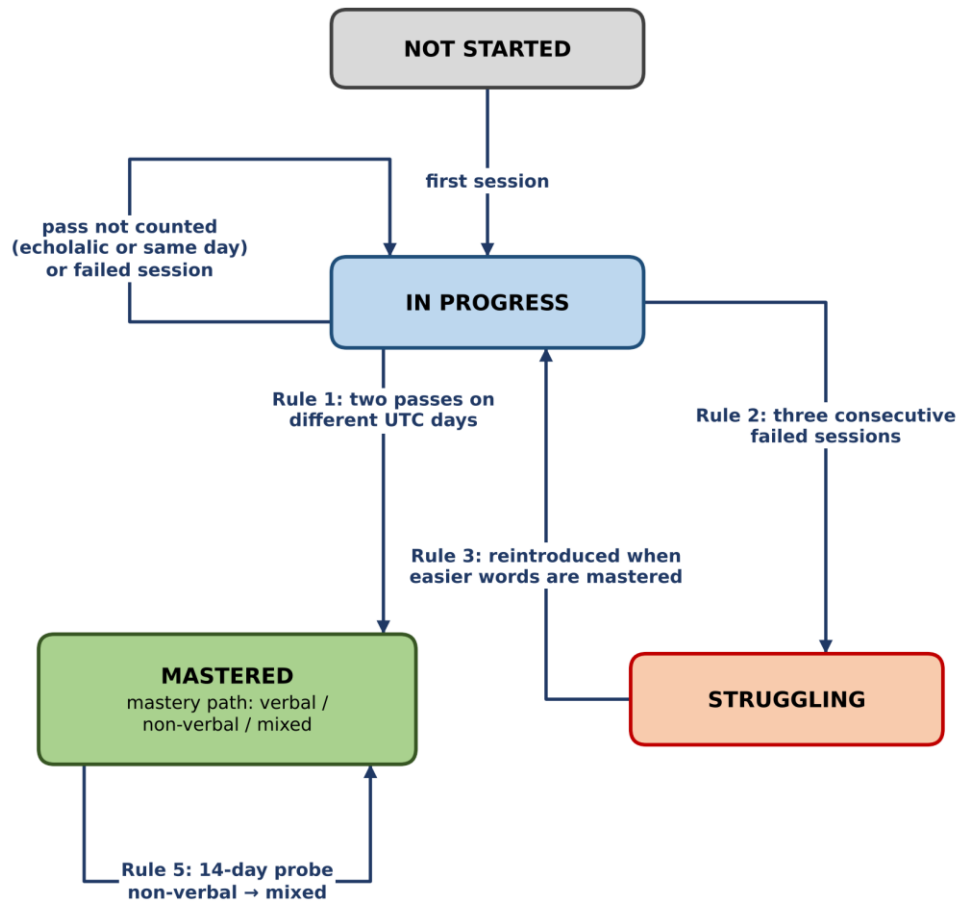


Figure 4.5: Word status states and transitions

4.4.4 RC1: Phoneme-aware speech assessment

RC1 is split between the backend and the microservice. The backend converts the recording, calls the speech service (British English, the short-command model, word timings) and normalises the transcript. It then applies the first layer, the accepted-form check: each word has a list of accepted full forms, scored 3, and key phrases, scored 2. Only transcripts that match neither are sent to /score-phoneme.

The microservice converts the transcript and the target into phoneme sequences with the CMU Pronouncing Dictionary, aligns the two sequences and classifies each difference against the Sinhala L1 transfer-error taxonomy in Table 4.5. The score is then set by the rules in Table 3.3, and the service returns the score, the phoneme accuracy and the main error class. The fallback, used only when the service does not

respond within two seconds, scores 3 for a token-sort similarity of 80 or more, 2 for 60 or more, 1 for a keyword match and 0 otherwise.

Table 4.5: Sinhala L1 transfer-error taxonomy used by RC1

Error class	Example	Reason for the error
TH-fronting	"thank" heard as "fank"	Sinhala has no dental fricative /θ/
TH-stopping	"thank" heard as "tank"	Sinhala has no dental fricative /θ/
DH-stopping	"the" heard as "de"	Sinhala has no voiced dental fricative /ð/
Final consonant deletion	"good" heard as "goo"	Sinhala words rarely end in a consonant
V/W confusion	"welcome" heard as "velcome"	Sinhala does not contrast /v/ and /w/
Vowel substitution	"sorry" with a different vowel	Different vowel inventories
R deletion	"morning" without /r/	Influence of non-rhotic Sri Lankan English
Cluster reduction	"brush" heard as "bush"	Sinhala has few consonant clusters
Voicing error	"bye" heard as "pie"	Different voicing contrasts

The examples in Table 4.5 are illustrative. The classes follow the difficulties reported for Sinhala-speaking learners of English, who replace the "th" sounds, confuse /w/ with /v/, devoice final consonants and simplify consonant clusters [12]. Two further difficulties reported there, /f/ replaced by /p/ and /z/ by /s/, are not separate classes in the current taxonomy and are reported as a limitation in Chapter 7. Scoring such features as partial rather than failed attempts is consistent with the view that Sri Lankan English has its own legitimate norms [24], [25]. A score of 2 with a flagged error class tells the teacher that the child produced the word with a known, expected transfer error, which is useful teaching information and is not treated as failure.

4.4.5 RC3: Echolalia detection

RC3 needs no extra model. It uses three timestamps: the end of the avatar's audio and the start of recording, both sent by the app, and the offset of the first recognised word, returned by the speech service. Response latency is the recording start plus the first-word offset minus the end of the avatar's audio. An attempt with a latency under 1,500 ms is flagged. When an attempt has no usable timestamps, the flag stays false and the latency is stored as missing.

Each child's echolalia rate is held as an exponential moving average (EMA) with $\alpha = 0.3$, updated after every attempt. When the rate is above 0.5, the app opens the microphone 3,000 ms after the avatar finishes on the next attempt. All three values (1,500 ms, 0.3 and 0.5) are named, tunable defaults in the code. The flag, the latency and the rate are stored with each attempt and shown to the teacher per word.

4.4.6 RC2: Learning-trajectory prediction and explanation

Figure 4.6 shows how a trajectory label is produced. The trajectory service first assembles the 13 features listed in Table 4.6 from the child's first session with the word, reading Phase 1 and Phase 2 data from the correct table for the word's category (Abilities words are stored separately). If the machine-learning flag is off, the label is "typical". Otherwise:

1. The features are sent to /predict-trajectory with a two-second timeout. The microservice one-hot encodes the categorical features, re-indexes the result to the 20 columns stored with the model, and returns the predicted class and the share of trees that voted for it.
2. If the vote share is at least the confidence threshold (provisionally 0.60), the Tier 2 label is used.
3. If the vote share is lower, the call fails or times out, or the attempt is non-verbal, the Tier 1 formula (Section 3.3.4) gives the label instead.

Table 4.6: Trajectory model features

Feature	Phase	Description
phase1_exposure_ratio	1	Exposures needed relative to the required count, recorded once at the Phase 1 gate; -1.0 for Abilities
speech_score	2	RC1 score, 0-3 (binary on the non-verbal pathway)
phoneme_accuracy	2	Proportion of phonemes correct, 0-1
phoneme_error_class	2	Main error class (categorical); the synthetic-trained model knows six values (none, other, TH-fronting, V/W confusion, final consonant deletion, vowel substitution), and any other class is encoded as all zeros
response_latency_ms_phase2	2	Time from the end of the avatar's audio to speech onset
echolalia_flag	2	RC3 flag (0 or 1)

Feature	Phase	Description
response_latency_ms_phase3	3	Time from showing the scenario to the child's selection
first_tap_correct	3	Whether the first tap was the correct picture
selection_change_count	3	Number of changed selections: 0, 1 or 2 or more
prompt_count	3	Number of prompts needed
difficulty	-	Word difficulty, 1-3
category	-	Greetings, Magic Words or Abilities (categorical)
phase1_applicable	-	Whether Phase 1 exposures apply to the word

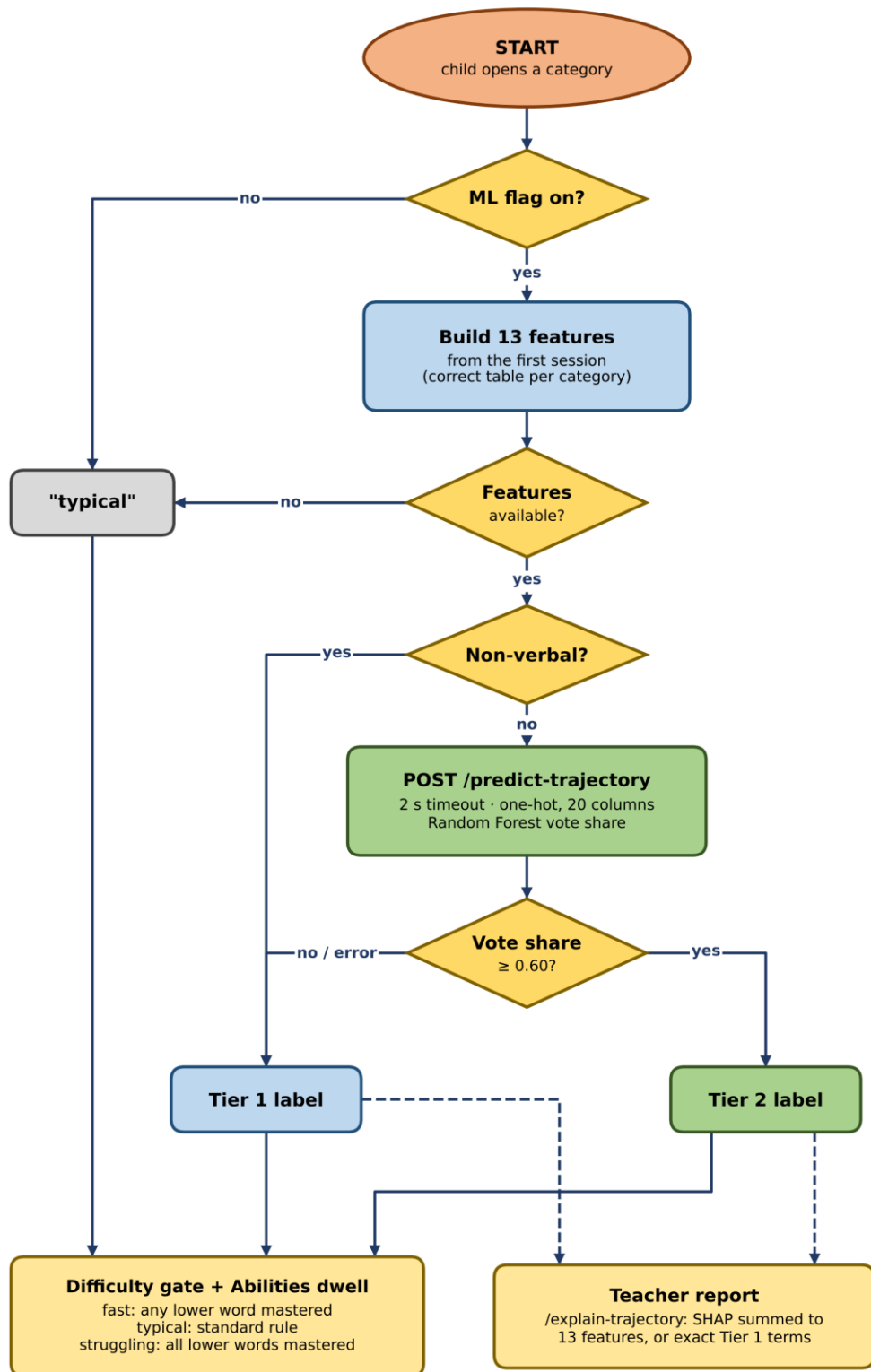


Figure 4.6: Trajectory prediction and explanation flow

The label is used in two places only: the difficulty gate (Section 4.4.3) and, for Abilities words, the familiarisation dwell, where for a "struggling" child the Next button waits for one full 1,800 ms animation cycle. The model file stores the trained forest, the label encoder, the column list and the scikit-learn version, so a mismatch between training and serving can be detected.

Explanations. For a Tier 2 label, /explain-trajectory computes SHAP values with a tree explainer built when the service starts, and sums the values of the encoded columns back to the 13 named features, so the teacher sees "speech score" rather than a list of encoded columns. For a Tier 1 label, the explanation is the exact contribution of each formula term. Explanation requests are separate from prediction requests, so a failed explanation never changes the label shown.

4.4.7 Teacher reports

The teacher reports turn the stored data into evidence the teacher can act on:

- **Trajectory report.** For each child, the predicted label for each word, which tier produced it, the confidence shown as a vote share, and the explanation. Every Tier 2 explanation carries the following note, visible without scrolling: "This is an early hint from a still-learning model, based on a very small amount of real data so far. It hasn't yet been shown to be reliable; right now it gets it right about as often as a guess would. Use your own observation of the session as the main guide, and treat this as one more thing to consider, not a conclusion." The note is a single named constant, so it can be updated in one place when the real-trained model replaces the current one.
- **Progress and history.** Status, mastery path, attempt history with dates, speech scores with flagged error classes, and echolalia frequency per word, with trend graphs.
- **Level 2 report.** Topic status and element scores, extended during integration to show sentence-by-sentence production, drag-and-drop and sentence-construction results.
- **Printing.** Reports can be printed for meetings with parents and other teachers.

[Needed for the final thesis: a figure of the teacher report screens (the trajectory report with the caveat note and an explanation; word progress and attempt history; the Level 2 report).

The screenshots were not available for this draft. When added, it becomes Figure 4.7, and the later Chapter 4 figures move up by one.]

4.4.8 Level 2 sentence construction

Level 2 was redesigned so that every activity traces to a page of the Grade 1-2 activity book used at the partner school (Table 4.7) and so that, as in Level 1, the child becomes familiar with each sentence before being asked to say it. Figure 4.8 shows the flow.

Table 4.7: Traceability of Level 2 activities to the Grade 1–2 activity book

Activity book page	Activity in the application
p1 "Myself - Draw yourself. Write your name."	"Draw yourself" canvas; the portrait then appears beside the child's sentences throughout Level 2
p2 "Make a badge. Write your name."	Name entry presented as making a name badge
p8 "How old are you? I'm 8."	Age picker using number characters; the app says the chosen age
p11 "Where do you live?" and the district map	Home-town picker from the Sri Lankan districts
p9 "Hi, I'm Kumar. I'm eight years old."	Sentence familiarisation steps for Self-Introduction
p12-14 "This is my friend...", "He is nine years old."	Describing a Friend: friend details and third-person sentences (He is / She is)
p15 "My pets" (cat, dog, cow, fish, parrot, rabbit) and p16 "Do you have a pet?"	Pet picker limited to these six animals, with an optional pet name
p17 "This is my pet. It is a cat."	Describing a Pet sentence patterns
p10 "What grade are you in?"	Not used for the child (out of scope for the target sentences); used only for the friend's grade

Flow. The child (with the teacher) chooses a topic, then completes the topic profile. A loading screen covers the time taken to generate the audio. A contrastive screen then shows how the same sentence is ordered in Sinhala (SOV) and in English (SVO), addressing a known source of word-order errors. For each sentence, the child goes through four steps: listen and watch the sentence build up word by word; build the sentence from shuffled word tiles; fill a single gap from three options; and say the sentence. Wrong tiles shake gently and return, and the correct answer for every gap always comes from the child's own profile, while the distractors come from fixed lists. The child then listens to the whole set together before the production phase. Figure 4.7 shows the flow and Figure 4.8 the main screens.

Sentence patterns. Table 4.8 lists the target sentences. Self-Introduction sentences are filled from the child's profile; the Describing a Friend sentences depend on which details were entered; and the Describing a Pet sentences are fixed for each animal, with the pet's name added when one was given.

Table 4.8: Level 2 target sentence patterns

Topic	Sentences
Self-Introduction	My name is [name]. / I am [age] years old. / I live in [home town]. / I am a [boy/girl]. / I like [activity].
Describing a Friend	My friend's name is [name]. / My friend is a [boy/girl]. / He (She) is [age] years old. / He (She) is in grade [grade]. / I like my friend because he (she) is [quality]. The last three appear only when those details are given.
Describing a Pet	Up to five fixed sentences for each of the six animals, starting from the activity-book pattern "This is my pet. It is a [animal]." The second sentence gives the pet's name and is left out when no name was given.

Production and scoring. For Self-Introduction, the child first tries the whole paragraph ("Hello! My name is ..."), and then each sentence in turn, prompted by its question (for example, "How old are you?"). Describing a Friend and Describing a Pet use only the sentence-by-sentence production. Every sentence has its own three-level triggers (the full sentence for 3, a shorter form for 2 and the key word for 1) and goes through the same assessment pipeline as Level 1. A sentence counts as a produced element when it scores 1 or more, and the session's element score is the number of elements produced. A topic is mastered with an element score of 4 or more in two sessions on different days, and is marked struggling with a score of 1 or less in three consecutive sessions. A child who does not speak can complete the topic by matching words, which is recorded as a non-verbal attempt.

Personalised audio. Sentence audio is generated with Azure TTS using Speech Synthesis Markup Language (SSML). The first design used two voices: a British English voice for English words and the eight anglicised districts, and a Sinhala voice chosen by the child's gender for the other 17 district names. The current design uses one high-definition voice (en-IN-Aarti) for the whole sentence at 60% of its normal rate, and marks Sinhala names (the 17 non-anglicised districts and the child's, friend's or pet's name in its teacher-confirmed Sinhala spelling) as Sinhala inside the same voice, so the sentence sounds like one speaker. The voice and rate can be

overridden by configuration without code changes. The child's name is given a suggested Sinhala spelling, which the teacher must confirm before it is used.

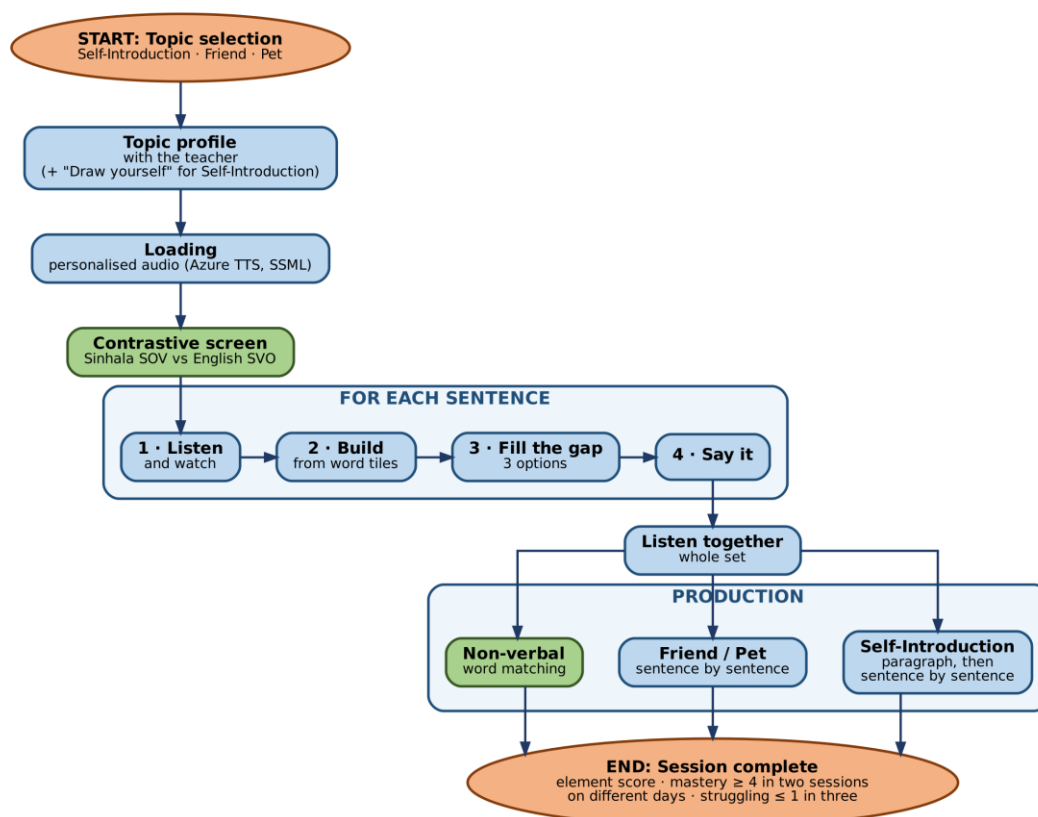


Figure 4.7: Level 2 flow

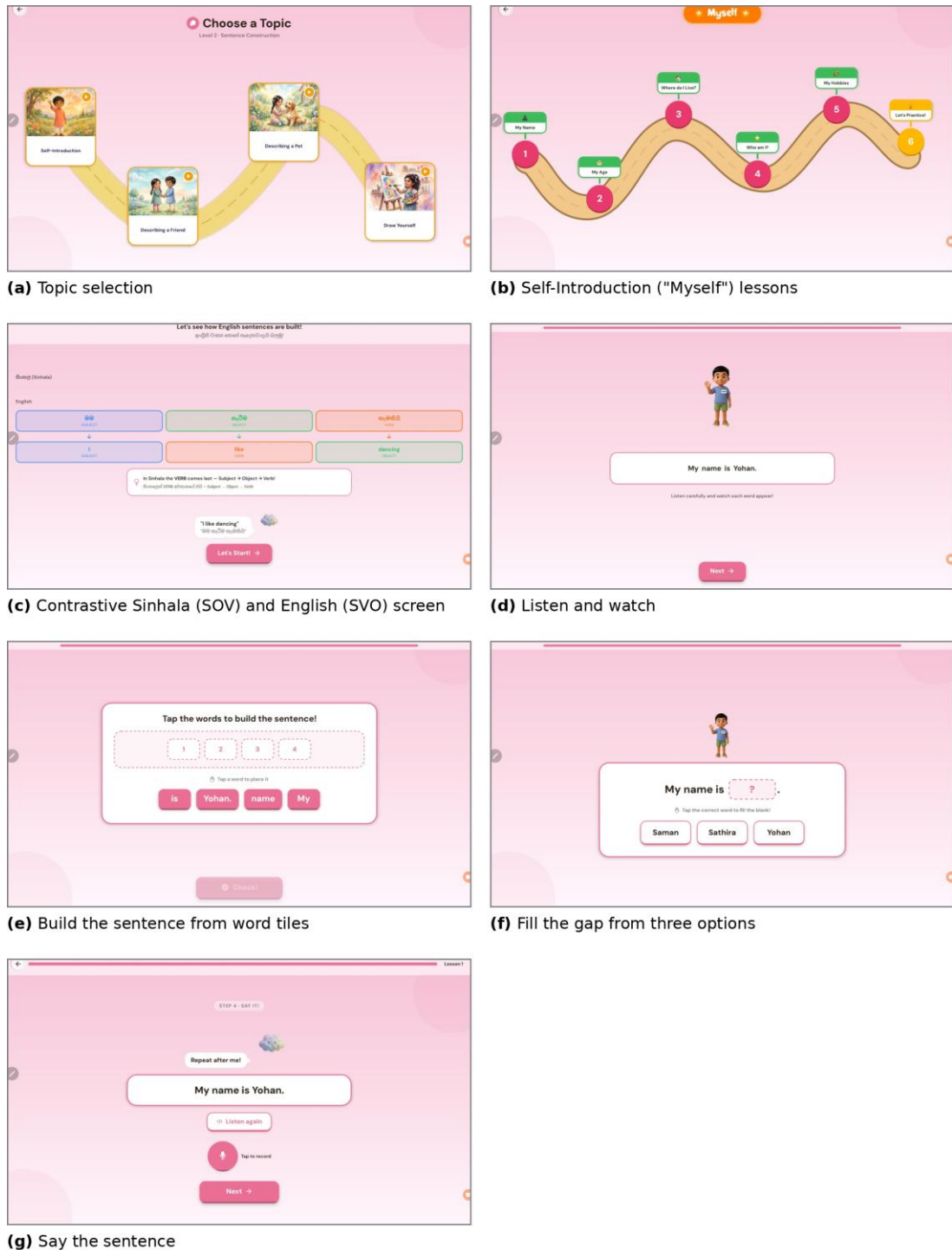


Figure 4.8: Level 2 screens (Self-Introduction)

4.4.9 Data design

The component's data is held in 12 tables of the shared database, linked to the platform's shared student table (Table 4.9). Figure 4.9 shows the entity-relationship (ER) diagram. Three principles guided the data design: personal details for Level 2

are kept in their own table, apart from interaction data; every attempt is stored as its own row with its timestamps, so analysis never depends on summary counters; and values that the model uses as features, such as the Phase 1 exposure ratio, are recorded once and never overwritten.

Table 4.9: Database tables owned by the component

Table	Contents
dialogue_words	The Level 1 words: category, difficulty, accepted forms, key phrases and prompt cue
dialogue_word_progress	One row per child and word: status, phase, pass and failure counters, mastery path, exposure snapshot, probe date
dialogue_word_attempts	Phase 2 attempts for Greetings and Magic Words: transcript, scores, phoneme results, timestamps, latency, echolalia flag
dialogue_phase3_attempts	Phase 3 scenario attempts: scenario, correctness, latency, first tap, selection changes, prompts
action_word_attempts	Abilities attempts across phases
dialogue_evaluation_attempts	Formative evaluation results
sentence_questionnaires	Level 2 profile for each child and topic, including the confirmed Sinhala spellings and the portrait strokes
level2_sessions	Level 2 sessions: topic, pathway and element scores
level2_sentence_attempts	Per-sentence familiarisation and construction results
level2_production_attempts	Spoken Level 2 attempts with their scores
level2_nonverbal_attempts	Non-verbal Level 2 attempts
level2_topic_progress	Status and counters for each child and topic

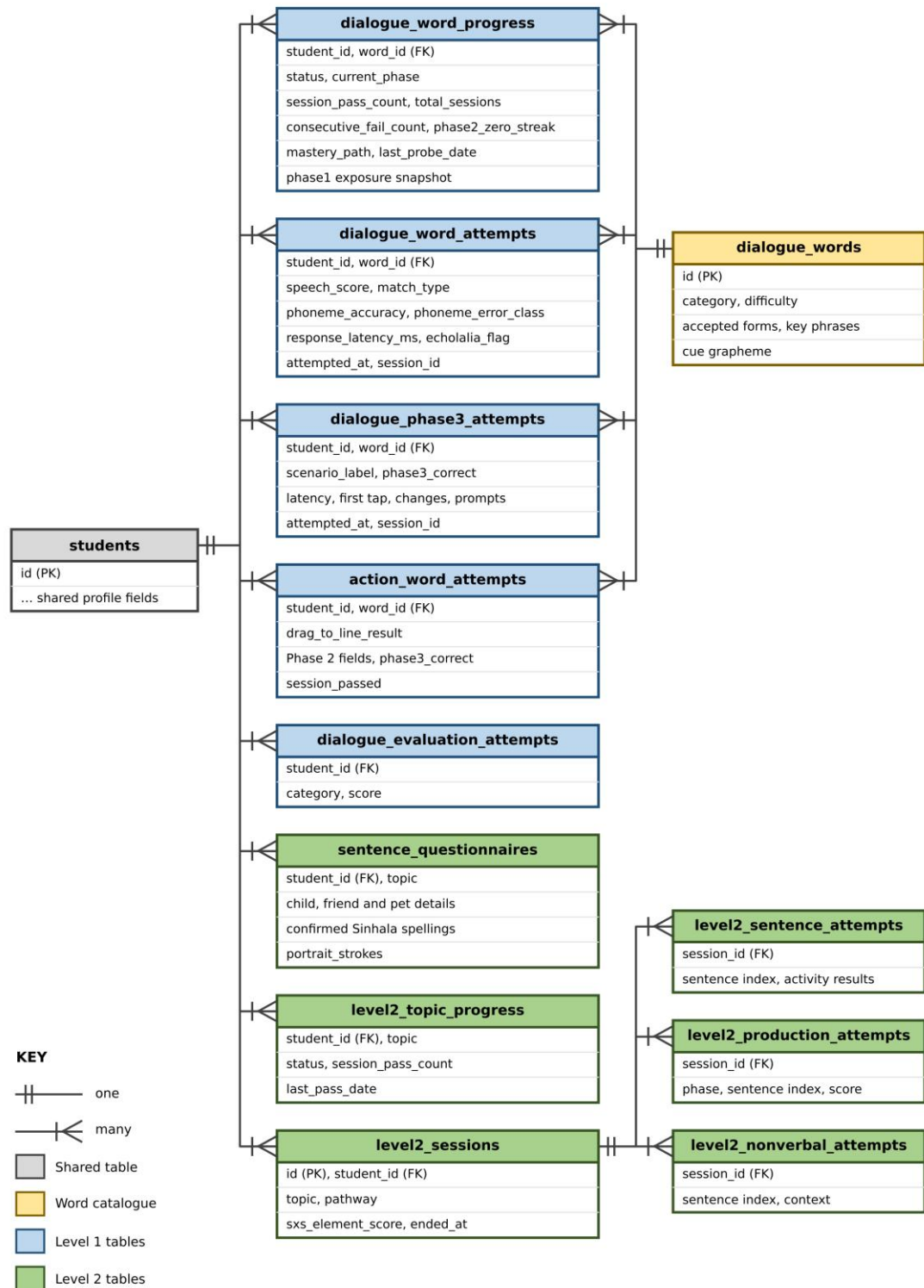


Figure 4.9: Entity-relationship diagram of the component's tables

4.4.10 Child-safety design principles

Five principles apply across every screen, and each is enforced in the design rather than left to the teacher:

1. **No failure state for the child.** No screen shows a score, a red mark or failure language; feedback is always graded and positive.
2. **No dead ends.** Every phase has a non-verbal route, and repeated difficulty leads to a gentler path, not a repeated demand.
3. **No loops.** Retries, rewatches and prompts are capped, and wait times shorten after refusals.
4. **Adult control.** The teacher can pause or exit a session at any time, and an adult check guards backward navigation.
5. **Models advise, rules decide.** No model output can change a child's recorded mastery, and every model has a deterministic fallback.

4.5 Standards and Professional Practice

The component was designed and built following these practices:

- **Requirements and traceability.** Every development task had a written specification with acceptance criteria, and every design decision has a numbered entry in the decision log that links to the tasks that implemented it.
- **Version control.** Each repository is under Git, with personal branches in the shared AURIVA repositories. Credentials and environment files are never committed.
- **Database change control.** Every schema change is a reviewed migration, run manually after review.
- **Testing.** Automated unit tests for the core logic and verification against acceptance criteria for every task (Chapter 5).
- **Data protection.** Children are identified only by codes in research data, access to the database is limited to the research group, the speech service is used with data logging disabled, and no audio is stored.

- **Responsible AI.** The model is kept behind a flag, confidence is reported honestly as a vote share, every prediction is explained, and the limits of the model are stated to the teacher.
- **Accessibility of the interface.** Large touch targets, minimal text, consistent layouts across the four components, and a non-verbal route through every activity.
- **Reproducibility.** Library versions are pinned, the random seed is fixed, and the model file records the library version used to train it (Section 3.7).

5 IMPLEMENTATION, INTEGRATION AND TESTING

This chapter describes how the design in Chapter 4 was built: the tools and technologies used (Section 5.1), the main implementation work and the problems solved during it (Section 5.2), how the component was integrated into AURIVA (Section 5.3), how it was verified (Section 5.4) and how it was deployed for the field study (Section 5.5).

5.1 Development Environment and Technology Stack

The component was developed in three repositories. Table 5.1 lists the main technologies used in each.

Table 5.1: Technology stack

Layer	Technologies	Purpose
Mobile app	React Native 0.81 with Expo 54; React Navigation; expo-audio and expo-av; react-native-svg; expo-print; sinhala-unicode-converter	Child and teacher screens; audio recording and playback; the "Draw yourself" canvas; printable reports; suggested Sinhala spellings
Backend	Node.js with Express 5; Sequelize 6 with PostgreSQL; jsonwebtoken; fluent-ffmpeg and ffmpeg-static; fuzzball; @google-cloud/speech; Jest 30	API and learning logic; database access; audio conversion; fallback scoring; speech-to-text; unit tests
Microservice	Python with Flask; pronouncing (CMU dictionary); pandas and NumPy; scikit-learn 1.6.1; joblib 1.4.2; SHAP	Phoneme scoring, trajectory prediction and explanations
External services	Google Cloud Speech-to-Text; Azure Text-to-Speech	Transcription; personalised Level 2 audio
Model training	Google Colab notebook	Algorithm comparison and model training

Development was done in Visual Studio Code on Windows, with Git for version control and Beekeeper Studio for inspecting the database. The development process (about 45 specified tasks, a decision log and AI-assisted implementation) is described in Section 3.1 and declared in Appendix B.

5.2 Component Implementation

This section covers the implementation work that was not straightforward and that shaped the final behaviour. The design itself is not repeated.

Audio pipeline. The app records in a compressed format (m4a), but the speech service needs uncompressed audio for reliable results with short clips. The backend therefore converts each recording to 16-bit linear WAV with ffmpeg before transcription; this was the only new backend dependency, and it was approved and logged as a decision. The record button is wrapped in a guarded recorder hook, so a second tap while recording is starting or stopping is ignored, which prevents the overlapping recordings seen in the classroom.

Learning logic. The dialogue service applies the mastery rules at the end of every session. Three implementation details mattered for data quality:

- the Phase 1 exposure ratio is written once, when the child passes the Phase 1 gate, into a separate snapshot field, because the live counter resets every session and could not be used as a feature;
- failed sessions and failed attempts are counted by two separate counters; when one counter was used for both it could exceed the number of sessions, which was one of the two faults that corrupted the pilot struggling labels;
- calendar days for Rule 1 are compared as UTC dates.

The prompt hierarchy is configured per word: each word stores a fixed cue that marks the letters to highlight, and the slowed replay plays the same recording at 60% speed. Wait-time escalation is computed from the day's attempt records, so it needs no extra storage, and any speech attempt resets it. The rewatch cap and the adult check before backward navigation are enforced in the app. The evaluation service unlocks a category's evaluation once three of its words are mastered, builds at most five word–picture pairs, and stores the score for the teacher only. The production probe writes only the probe date and the mastery path, never the word's status. Sinhala instructions are human-recorded audio files played through a single instruction-audio helper that checks the child's Sinhala setting, so every new screen received the Sinhala layer without extra work.

Speech assessment. The accepted-form check, the call to the phoneme service and the fuzzy-matching fallback are in one backend service. Every call to the microservice has a two-second timeout, and the use of the fallback is written to the

log as a warning. This warning is what made it possible to discover that RC1 had been silently replaced by the fallback during the early pilot.

Transcription uses synchronous recognition with British English, the short-command model and word time offsets. Transcripts are lower-cased and stripped of punctuation before matching. In the microservice, the transcript and the target are converted to CMU phoneme sequences and aligned with a sequence-matching algorithm, and each mismatched segment is assigned a taxonomy class. The fallback uses a token-sort similarity ratio, which ignores word order.

Microservice. The Flask service loads the model file once at start-up. The file contains the forest, the label encoder, the column list and the scikit-learn version. The SHAP explainer is also built once at start-up. Each prediction request is encoded as a single row and re-indexed to the stored columns, and explanations are summed from the 20 encoded columns back to the 13 named features.

Data and model pipeline. The microservice repository holds the scripts behind every dataset used in this thesis: an extractor that builds the trajectory dataset from the database (keeping only the three Level 1 categories, using the first Phase 3 scenario of the first session, assigning labels by the mapping in Table 3.2 and excluding words still in progress); a script that checks struggling rows for the counter fault; a diagnostic script that reports missing values per feature; and the synthetic data generator with its variant manifest. Models are trained in a Google Colab notebook and exported with the forest, column list, label encoder and library version. The faulty calibrated model is kept under a separate name, marked as broken, as evidence for the audit.

Level 2. The sentence builders create each topic's sentences from the child's profile, and each sentence gets its own three-level speech triggers, built from the same text. The "Draw yourself" canvas stores strokes as data and redraws them wherever the portrait appears. An SSML builder marks Sinhala names inside the English sentence before the audio is generated (Section 4.4.8). Suggested Sinhala spellings of names were correct for 15 of 19 real names in testing, so a spelling is saved only after a teacher confirms it.

The text-to-speech service generates each sentence's audio when the session starts, so the child never waits between sentences. The district pronunciation map and the fixed pet sentences are stored as data files, so they can be corrected without code changes. The schema changes for Level 2 added the friend and pet topics, their profile fields, and a column for the portrait strokes.

Teacher reports. Report services gather attempts, progress and predictions for each child. The Tier 2 caveat note is a single named constant, and reports can be printed with expo-print. The component's own API endpoints are documented in the backend's interactive API documentation (Swagger).

5.3 Integration with the Other Components

The four AURIVA components were integrated into one application by the group leader, who also deployed it. The app and backend code of each component were merged into the group's shared app and backend repositories, where each member worked on a personal branch (this component's backend branch is dev/maryse). The components share authentication, student profiles and one PostgreSQL database. Each member's Python microservice remains a separate repository and service; this component's is auriva-microservice.

This component's code was taken at its latest commit, and integration required no changes to its learning logic, assessment or models. Two frontend changes were made: the screens were restyled so that all four components look the same, and the Level 2 teacher report was extended to show sentence-by-sentence production, drag-and-drop and sentence-construction results. Because all components write to one database, each component keeps its own tables, and every schema change is applied as a reviewed migration. Access to the database is limited to the four group members. The Google Cloud service-account key was added to the integrated deployment by hand and was never committed to a shared repository.

Authentication and general security are provided by the shared AURIVA backend, not by this component: token-based login, role checks on routes, hashed passwords, standard security headers, request rate limiting, input validation, and request and error logging. This component's routes use the shared authentication and role checks.

5.4 Verification and Software Testing

Acceptance verification. Every task was checked against its written acceptance criteria before it was approved. Backend changes were first checked with syntax and module-loading checks, and anything the child or teacher would see was then tested live on the device.

Automated tests. The backend has 114 Jest unit tests in four suites (Table 5.2). They cover the parts of the logic where a silent error would change what a teacher sees or what a child is offered next. All 114 tests pass on the component's backend code at the testing-build version (four suites, run in 1.4 seconds).

Table 5.2: Automated test suites

Suite	Tests	What it covers
tier1Scorer	15	Formula terms, weight redistribution when a term is missing, binary non-verbal scores, class thresholds
trajectoryService	57	Feature assembly for each category, the flag-off default, the confidence gate, fallback on timeout or error, non-verbal routing, explanation handling
dialogueAnalyticsService	16	Level 1 report figures
level2AnalyticsService	26	Level 2 report figures, including the integration additions

The microservice has no automated test suite. It was verified by live start-up checks (the model loads and its version matches), by direct requests to each endpoint, and by a sweep of 6,482 generated inputs, which confirmed that adding explanations changed no predicted labels.

Audits. Three structured audits were carried out alongside normal testing: a code audit (16 August 2026), an audit of the trajectory model's calibration (19 August 2026) and an audit of the training labels (6 October 2026). Table 5.3 lists the main defects found by testing, audits and pilot use, and how each was resolved. The open items are discussed as limitations in Chapter 7.

Table 5.3: Main defects found and their resolution

Defect	How it was found	Resolution
Phoneme service not running in early pilot sessions, so all scoring silently fell back to fuzzy matching	Fallback warnings in the logs; no phoneme values in the pilot data	Pre-session checklist and smoke test; confirmed working from 25 July 2026

Defect	How it was found	Resolution
Pilot tablet running an old build, so Phase 3 telemetry was missing	Telemetry gap in the data	Rebuild before every session; a frozen testing build for the field study
Failed attempts counted as failed sessions	Label audit	Separate counters (fixed 6 August 2026)
Correct non-verbal sessions scored as failures	Label audit	Non-verbal threshold of 1 (fixed 6 August 2026)
Missing response timestamps for 42 Phase 2 attempts	Data extraction	Missing latency stored as missing, not zero
Duplicate and orphan rows in one category	Data extraction	Cause fixed; remaining orphans removed during extraction
Calibration fitted on nine rows, making the model predict "fast" for every input	Calibration audit	Calibration removed; confidence shown as a vote share
Evaluation script encoding all rows together, which zero-filled the first_tap_correct column	Cross-check against the audit's own figures	Rows encoded one at a time, as in production
Wait-time escalation inactive for Abilities	Code audit	Open (Chapter 7)
Rule 4 reduction branch unreachable	Code review	Open (Chapter 7)

The evaluation-script defect shows why the evaluation must mirror production. A single missing value in first_tap_correct changed the data type of that column for the whole batch, so the column was wrongly encoded and then filled with zeros for every row. Production never meets this problem because it encodes one request at a time. After the fix, the script's results matched the audit's independently reported figures exactly.

The label audit also checked the database for internal consistency. Across 104 progress records, no word had more passed sessions than total sessions; no record created after the counter fix had more failures than sessions; and no word was mastered with fewer than two passes. Finally, a start-up script launches the backend and the microservice and lists the manual checks to complete before each session: rebuild the app on the tablet, then complete one Phase 2 attempt and one Phase 3 scenario while watching the service logs for fallback warnings.

5.5 Deployment and Operation

The field study uses the deployed AURIVA testing build (auriva-db5c387) described in Section 3.3.7. It was built from the latest commit of every repository, was changed

during testing only to deploy the single-voice Level 2 audio, and is installed on an Amazon Fire HD 10 (11th generation) and a Samsung tablet. The backend and microservices run on Azure. The deployed model is the uncalibrated, synthetic-trained Random Forest, with the machine-learning path switched on for the supervised sessions and a Tier 2 confidence threshold of 0.60.

Before each testing day, the device's build is confirmed, the backend and the phoneme service are confirmed to be reachable, one Phase 2 attempt and one Phase 3 scenario are completed as a smoke test, and the database is checked to confirm that scores, latencies and phase records were stored. The single-voice Level 2 audio (Section 4.4.8) was deployed during the field study. After the field study, the trajectory model will be retrained on real data only (Section 3.3.4). Because the column list and label encoder are stored with the model, the retrained model replaces the current file in the microservice without any code changes.

6 RESULTS AND VALIDATION

This chapter reports results only; they are interpreted in Chapter 7. Some results depend on field study sessions that had not finished when this draft was written; each such section states that its results are pending and when they will be available.

6.1 Research Results

6.1.1 RC2: Algorithm comparison on synthetic data

Five algorithms were compared on the 400-row synthetic dataset (100 simulated children; 12% fast, 32% typical, 56% struggling) using leave-one-child-out cross-validation, so that each of the 100 folds held out all four rows of one simulated child. All models used balanced class weights except the multilayer perceptron, which does not support them. Table 6.1 gives the pooled out-of-fold results from the training notebook.

Table 6.1: Algorithm comparison on synthetic data (n = 400, leave-one-child-out cross-validation)

Algorithm	Accuracy	Balanced accuracy	Macro-F1
Random Forest (300 trees)	0.915	0.892	0.897
Multilayer perceptron	0.903	0.873	0.881
Logistic Regression	0.903	0.895	0.880
XGBoost	0.895	0.885	0.879
Support vector machine (RBF kernel)	0.893	0.882	0.874

Random Forest had the highest accuracy and macro-F1; the other four models were close behind, and Logistic Regression had a marginally higher balanced accuracy. The Random Forest recalled 85% of fast, 96% of struggling and 87% of typical rows. A Keras neural network was also set up in the notebook, but its run has no saved output and it is not reported. The deployed model was then trained on the 2,000 pooled rows of the five domain-randomised variants, producing 20 encoded columns. The ten most important features in that forest were `prompt_count` (0.199), `speech_score` (0.141), `phoneme_accuracy` (0.121), `first_tap_correct` (0.096), `selection_change_count` (0.085), `response_latency_ms_phase2` (0.082),

phase1_exposure_ratio (0.078), response_latency_ms_phase3 (0.072), echolalia_flag (0.044) and difficulty (0.019).

As a quick check, the Random Forest from this comparison was also scored on the 27 real pilot rows, with missing values filled from synthetic averages: accuracy 0.481, balanced accuracy 0.509 and macro-F1 0.408, with no struggling row classified correctly. Because some inputs were filled in rather than real, this check is reported only as supporting evidence for Section 6.1.3.

6.1.2 RC2: Calibration audit

The audit of 19 August 2026 found that the deployed model's probability calibration had been fitted on nine real rows, only one of them struggling. These were the only verbal-pathway rows with complete features, from three children. In the notebook's leave-one-child-out check of this step, the calibrated model classified all six fast rows correctly but neither typical row nor the single struggling row (accuracy 0.666, 90% bootstrap interval 0.439-0.889; macro-F1 0.341). The fitted curve for the struggling class was inverted, so the calibrated model predicted "fast" for essentially every input, whatever its features. The calibration was removed, and all results below use the uncalibrated Random Forest.

6.1.3 RC2: Synthetic-trained model on the real pilot set

The uncalibrated synthetic-trained model and the Tier 1 formula were scored on the 27-row real pilot set (8 fast, 9 typical and 10 struggling rows from 6 children), encoding each row on its own as in production. Table 6.2 gives the results on the full set, on the verbal-pathway rows (the only rows Tier 2 handles in deployment) and on the complete-case verbal rows with no missing telemetry.

Table 6.2: Synthetic-trained Tier 2 and Tier 1 on the real pilot set

Subset	n	Tier 2 accuracy	Tier 2 macro-F1	Tier 1 accuracy	Tier 1 macro-F1
All rows	27	0.444	0.364	0.407	0.333
Verbal pathway	15	0.600	0.422	0.667	0.468
Complete-case verbal	9	0.778	0.508	-	-

In the complete-case subset, Tier 2 classified all 6 fast rows correctly and the single struggling row wrongly, giving a balanced accuracy of 0.50. Neither tier classified a

struggling row correctly in any subset. Tier 1 was ahead of Tier 2 on the verbal-pathway rows and narrowly behind it on the full set. In this evaluation, the Tier 1 latency ceiling was set from the 90th percentile of the real Phase 3 latencies rather than the deployed 8,000 ms, so the Tier 1 figures approximate, but do not exactly reproduce, the deployed formula; Tier 1 is re-scored with the deployed setting in Section 6.1.5.

6.1.4 Label audit

The audit of 6 October 2026 checked whether the real labels could be trusted. In the progress table, 11 of the 12 word records with struggling status had more consecutive failures than total sessions, which the rules make impossible. All 10 struggling rows in the evaluation set had this fault, and all 10 were Abilities words from 4 children; none of the 21 fast and typical rows had it. Re-applying the corrected rules recovered none of the struggling labels. Since the fix on 6 August 2026, about 14 trustworthy rows exist, from 6 children, and none is labelled struggling. The struggling rows in Table 6.2 were therefore produced by the faulty counter: the struggling results in Section 6.1.3 do not measure real struggling behaviour, and only the fast and typical results rest on sound labels.

6.1.5 RC2: Final model trained on real data

The final model was evaluated with leave-one-child-out cross-validation on the 87-row real dataset described in Section 3.2.4: 26 learner accounts, 31 fast, 30 typical and 26 struggling rows, and therefore 26 folds. All of these rows, including the 26 struggling rows, were created after the label fixes of 6 August 2026, unlike the faulty pilot labels in Section 6.1.4. Because struggling rows came from 17 accounts, the model has three classes. The Tier 1 formula, with its deployed settings including the 8,000 ms latency ceiling, and the synthetic-trained Random Forest, as deployed and without calibration, were scored on the same rows. Neither was trained on these rows, so no folds are needed for them. Every row was encoded on its own, as in production. Results are reported for the 57 verbal rows from 23 accounts, the only rows Tier 2 handles in the product, and for all 87 rows. Table 6.3 gives the results, with 90% confidence intervals for balanced accuracy from a child-level bootstrap with 2,000 resamples.

Table 6.3: Tier 1, synthetic-trained and real-trained models on the real dataset (RF: Random Forest; the real-trained RF is evaluated leave-one-child-out)

Model	n	Accuracy	Balanced accuracy (90% CI)	Macro-F1	Recall fast	Recall typical	Recall struggling
Verbal rows only							
Tier 1 formula	57	0.509	0.498 (0.416-0.573)	0.478	0.846	0.091	0.556
Synthetic-trained RF	57	0.719	0.723 (0.633-0.825)	0.735	0.846	0.545	0.778
Real-trained RF	57	0.702	0.695 (0.598-0.797)	0.682	0.692	0.727	0.667
All rows							
Tier 1 formula	87	0.391	0.377 (0.324-0.428)	0.313	0.871	0.067	0.192
Synthetic-trained RF	87	0.655	0.644 (0.565-0.748)	0.644	0.871	0.600	0.462
Real-trained RF	87	0.736	0.739 (0.671-0.811)	0.736	0.677	0.733	0.808

On the verbal rows, both Random Forests were well ahead of Tier 1, with balanced accuracies of 0.723 for the synthetic-trained and 0.695 for the real-trained model against 0.498, and neither forest's confidence interval overlapped Tier 1's. The two forests' intervals overlapped almost completely, so this data cannot separate them. Tier 1 almost never predicted "typical" (recall 0.091 on the verbal rows and 0.067 on all rows).

On all rows, the real-trained model scored highest: accuracy 0.736, balanced accuracy 0.739 and macro-F1 0.736, against balanced accuracies of 0.644 for the synthetic-trained model and 0.377 for Tier 1. A majority-class baseline run inside the same folds reached a balanced accuracy of 0.173, and in a permutation test none of 100 runs with shuffled labels matched the real result ($p \approx 0.01$). Most of the real-trained model's lead on all rows comes from the non-verbal rows, which it never sees in the product: 17 of the 26 struggling rows are non-verbal, and the one-line rule "non-verbal means struggling" would catch those 17 while wrongly flagging 13 other rows. Two further checks address this shortcut. A two-class version trained and tested only on the verbal fast and typical rows (48 rows) reached a balanced accuracy of 0.729 against a baseline of 0.5 (permutation $p = 0.02$). With incomplete rows

dropped instead of filled, the three-class model reached a balanced accuracy of 0.768 (90% CI 0.639-0.896).

The most important features of the final forest were `phase1_exposure_ratio` (0.199), `response_latency_ms_phase3` (0.192), `speech_score` (0.109), `category_abilities` (0.101) and `phase1_applicable` (0.081). The category features rank highly because 19 of the 26 struggling rows are Abilities words. Importances from 87 rows are unstable, so only the ranking should be read.

Confidence threshold. The rule in Section 3.3.4 was applied to the verbal rows. At every threshold from 0.34 to 1.00, the real-trained model's balanced accuracy on the rows it would handle was at least Tier 1's on the same rows: for example, 0.695 against 0.498 at 0.34, where it handles all 57 rows, and 0.834 against 0.556 at 0.60, where it handles 31. The rule therefore gives 0.34, the lowest value tested. With three classes, the winning class always has at least a third of the votes, so at this threshold the low-confidence fallback is never triggered; Tier 1 is then used only for non-verbal attempts, when the machine-learning flag is off, or when the call fails or times out. Raising the threshold made the system as a whole worse on the verbal rows (balanced accuracy 0.641 at 0.60), because the rows passed to Tier 1 were ones Tier 2 classified better.

6.1.6 RC1: Speech assessment agreement

This evaluation is pending. It compares the system's decision with the observer's on matched attempts, using the accepted or needs-more-practice mapping described in Section 3.4, together with agreement on the error type and how often the fallback was used. The observer sheets are still being collected during the field study, which is planned to end in the third week of October 2026, and the results will be reported in the final thesis.

6.1.7 RC3: Echolalia detection

This evaluation is pending. It measures the precision and recall of the echolalia flag against the echoing recorded by the observer under the ground-truth rule in Section 3.4, the distribution of response latencies, and how often the three-second

microphone delay was triggered. It depends on the same matched observer records as Section 6.1.6 and will be reported in the final thesis.

6.2 Product and Technical Results

Table 6.4 reports the technical results that were available when this draft was written.

Table 6.4: Product and technical results available for this draft

Measure	Result	Source
Automated unit tests passing	114 of 114	Jest run on the testing-build code (Section 5.4)
Predicted labels changed by adding explanations	0 of 6,482 generated inputs	Explanation sweep (Section 5.4)

The measures taken from the field study logs are pending: the share of sessions completed without a technical failure, the share of Phase 2 attempts scored by the phoneme service rather than the fallback, the median and longest time from the end of a recording to feedback, the number of trajectory predictions made by each tier, and the number of attempts with missing timestamps or phase records. They will be calculated from the system logs and the database once the field study ends.

6.3 User or Stakeholder Validation

The user validation is pending. Ten children and about six teachers at three schools in the Colombo district have taken part in the field study so far (Section 3.2.6). The final number of children (verbal and non-verbal or minimally verbal), teachers and sessions at each school, and any sessions stopped early, will be reported when the field study ends.

6.3.1 Teacher task usability

The task usability survey (Appendix A) is pending. It had not been conducted when this draft was written and will be administered in the remaining field sessions. The final thesis will report, for each of its 14 tasks, whether the teacher completed it unaided, with help or not at all, and the median and range of the ease ratings; the ratings of the nine statements on supporting the child; and each teacher's System Usability Scale score, with the mean and range, compared with the commonly cited average of 68.

6.3.2 Teacher feedback on the component's judgements

The feedback questionnaire is pending and will be given in the remaining field sessions. It will provide the teachers' agreement with the app's speech and echolalia judgements; the agreement between each teacher's own judgement of a child's learning pace and the app's label, bearing in mind that the deployed model is the synthetic-trained one (Section 3.3.7); ratings of how clear and useful the explanations were; and ratings of suitability for the children.

6.3.3 Child engagement and observations

These results are pending. The observer sheets record engagement, the points where engagement dropped, distress events, interventions and session completion, and the observer notes and open-ended answers will be grouped into themes. They will be analysed once all sessions are complete, at the end of the field study, which is planned to end in the third week of October 2026.

6.4 Integration Results

The component ran inside the deployed AURIVA testing build throughout the field study, using the shared login, the shared student profiles and the shared database. [Pending: any problems observed when moving between AURIVA components, whether the shared profile details were correctly available in the component, and whether the integrated Level 2 report showed the sentence-by-sentence, drag-and-drop and sentence-construction results correctly.]

6.5 Summary of Findings

Table 6.5 summarizes the findings for each research question that could be answered for this draft.

Table 6.5: Summary of findings by research question

Question	Finding
RQ1 (RC1)	Pending: depends on the observer agreement in Section 6.1.6.
RQ2 (RC2)	Yes, for the attempts the model handles: on the verbal rows of the real dataset, both Random Forests predicted learning pace clearly better than the Tier 1 formula (balanced accuracy 0.695 for the real-trained and 0.723 for the synthetic-trained model, against 0.498). Training on real data did not improve on the synthetic-trained model for these rows. The earlier pilot struggling labels were faulty (Section 6.1.4), and the real dataset includes test accounts (Section 3.2.4).

Question	Finding
RQ3 (RC3)	Pending: depends on the precision and recall in Section 6.1.7.
RQ4 (usability)	Pending: depends on the teacher survey and questionnaire in Section 6.3.

7 DISCUSSION

This chapter interprets the results reported in Chapter 6; it does not repeat them. It explains why the results came out as they did, compares them with the literature and existing solutions, states the research and product contributions, sets out the limitations, and draws out the implications for teachers, researchers and developers. Interpretations that depend on field study data are marked as pending.

7.1 Interpretation of Findings

RC2: what the synthetic and real data showed. The Random Forest scored 0.915 accuracy on synthetic data (Table 6.1) but only 0.444 on the 27-row pilot set (Table 6.2). Part of this gap is circularity. The synthetic generator assigned its labels with the same mastery rules and the same assumptions about which features matter, so high synthetic accuracy mostly showed that the model had learned the generator's assumptions back. The feature importances illustrate this: the synthetic forest's top three features (prompt count, speech score and phoneme accuracy) are the three terms the Tier 1 formula weights most heavily. Part of the gap, however, came from the pilot labels, which were later found to be faulty (Section 6.1.4). On the larger post-fix real dataset, the same synthetic-trained model reached a balanced accuracy of 0.723 on the verbal rows, close to the model trained on real data (0.695), and both were well above Tier 1 (0.498; Table 6.3). Synthetic pretraining therefore did not fail as badly as the pilot evaluation suggested. What failed was treating synthetic accuracy as a measure of real performance, and trusting a small set of unaudited real labels.

Why the real-trained model was not better on verbal rows. The real dataset is small, with 57 verbal rows from 23 accounts and only 9 struggling verbal rows, and the forest is deliberately constrained, so there was little for it to learn beyond what the generator's assumptions already encoded. Its higher score on all rows comes mainly from the non-verbal rows, where most struggling labels sit: it partly learned that non-verbal and Abilities rows tend to be struggling, which is true of this data but irrelevant in deployment, where non-verbal attempts go to Tier 1. The test-account

rows may also have distorted the results in either direction. The two-class check on the verbal fast and typical rows (0.729 against a baseline of 0.5) shows that the model does learn something beyond that shortcut.

What the results support is the design around the model. A learned model clearly beat the literature-grounded formula on real data, so the second tier adds value; the Tier 1 fallback remains necessary for non-verbal attempts and failures; and the flag, the caveat shown to teachers and the vote-share confidence remain appropriate for a model trained on so few children. The confidence gate is kept even though the threshold rule currently sets it to its lowest value (Section 6.1.5), so that it can be raised if a later model performs worse. Tier 1's weakness, especially its near-absence of "typical" predictions, suggests that its placeholder weights and cut-offs should be revisited. It is a low bar, so beating it shows that the model is useful, not that it is reliable.

The calibration failure. Fitting a probability calibration on nine real rows, one of them struggling, made the model predict "fast" for every input. The lesson is general: methods designed to correct a model with a little data can do serious harm when "a little" means a handful of rows. Reporting confidence as a vote share rather than a calibrated probability describes more honestly what the model knows.

The label audit. All ten pilot "struggling" labels came from two counting faults rather than from children's behaviour, so the pilot evaluation could say something meaningful only about the fast and typical classes. In a small special education dataset, a single fault in how data is recorded can outweigh any choice of algorithm. This is why the final training procedure was fixed in advance, filters rows by date, is evaluated by leaving one child out at a time, and falls back to two classes if too few real struggling examples exist.

RC1 and RC3. These results are pending (Sections 6.1.6 and 6.1.7), so only the questions they will answer can be set out here. RC1 is intended to give partial credit for the transfer errors that fuzzy matching accepts as correct, so the key question is where the system and the human rater disagree: disagreement on substitutions such as "th" to /t/ would point to the taxonomy, whereas disagreement on clear

productions would point to speech-to-text errors. For RC3, the question is whether the 1,500 ms threshold separates observed echoes from genuine responses, and whether the latency distribution supports a single threshold or one that depends on the child.

Product and usability. The usability results are pending (Section 6.3). They will be interpreted in the final thesis, in particular whether teachers could complete the tasks unaided and how the SUS score compares with the average of 68.

7.2 Comparison with Literature Baselines and Existing Solutions

Speech assessment. Fuzzy matching correlates closely with human intelligibility scores on transcripts [11], but it judges spelling rather than sound, so its errors hide exactly the transfer patterns described for Sinhala-speaking learners [12]. Tran et al. improved recognition of non-native child speech from 62% to 80% by mapping transcripts to the nearest curriculum sentence [10]; that recovers what the child meant but erases the pronunciation evidence a teacher needs. RC1 makes the opposite trade-off. It keeps the error and classifies it, which gives the teacher more useful feedback, at the cost of depending on dictionary coverage and a correct transcript; the fallback keeps sessions running when that dependency fails. Earlier systems for autistic learners switched speech recognition off altogether [28], [29]. How RC1's agreement compares with the proposal's targets of 75–80% will be discussed when the field results are available.

Echolalia. The only automatic echolalia detector reviewed counts repeated word pairs in transcripts and was built for diagnosis [32]. RC3 works during learning, uses timing rather than text, and does not penalise the child, in line with recent reviews that treat echolalia as functional and advise against suppressing it [13], [19]. The trade-off is that timing alone cannot distinguish a fast genuine answer from a fast echo; this is acceptable because a false flag only withholds mastery credit for one session and never penalises the child, but it would be problematic if flags were used to judge the child.

Adaptation. Adaptive tutors such as Schodde et al. choose tasks with knowledge tracing and adapt reliably once enough data exists [14], [33]. This component has far

less data, so it pairs a small model with a literature-grounded formula and explains both. On the 27-row pilot set the formula was the safer choice (Table 6.2), but on the larger real dataset both learned models were clearly better on the rows they handle (Table 6.3). This supports giving the model priority once it has been trained and checked on real data, with the formula as a fallback rather than the default.

LLM dialogue. LLM systems produce richer conversation [15], [31] but have not been evaluated with autistic children in real settings or need a human to check every turn [16]. The test rounds that led to the removal of Level 3 agree: predictable, word-level practice suited these children better than generated conversation.

Proposal targets. The comparison with the proposal's targets (speech assessment accuracy of 75%, task completion of 70% and teacher satisfaction of 4.0 out of 5, used as reference points, Section 3.4) is pending, because all three depend on the field study.

7.3 Research Contribution

The research makes four contributions:

1. **Contextual word teaching as an assessed step.** Phase 3 turns knowing when to use a word into a measurable learning objective for autistic second-language learners, which none of the reviewed systems does (Table 2.2).
2. **Transfer-aware speech assessment (RC1).** A Sinhala first-language error taxonomy gives partial credit to expected transfer errors and flags them for the teacher.
3. **Latency-based handling of echolalia during learning (RC3).** RC3 turns a definition that the literature leaves open into a measurable rule and handles echoes without penalty, in line with current guidance.
4. **An explainable two-tier trajectory model, with methodological evidence.** The two-tier design, in which a learned model beat a literature-grounded formula on real children's data (Table 6.3). The research also documents how such a model can mislead when data is scarce: synthetic accuracy overstated real performance, a small-sample calibration made the model predict a single class, and faulty labels produced a misleading pilot result. Each problem was found by auditing,

and the final evaluation was run child by child, with baselines and confidence intervals.

Contribution statement. This research contributes an adaptive dialogue learning component for autistic Sinhala-speaking learners of English by demonstrating transfer-aware speech assessment, latency-based echolalia handling and an explainable two-tier trajectory model, using a deployed application evaluated computationally and in a supervised field study under real classroom conditions at three schools. The results show that, on real data, a Random Forest predicted learning pace better than a literature-grounded formula on the attempts it handles, although training on real data did not yet improve on a synthetic-trained model [Pending: and the field agreement of RC1 and RC3], while the main limitation is the small number of children, with test-account rows in the training data.

7.4 Product Contribution

The component is a working part of a deployed, integrated application in use with children and teachers at three schools. It provides 30 curriculum-traced Level 1 words and three Level 2 topics, a non-verbal pathway through every activity, personalized audio that pronounces Sri Lankan names correctly, and teacher reports that explain predictions and state their limits. Its engineering choices make it suitable for real school use: a deterministic fallback for every model and external service, a flag controlling the machine-learning path, 114 passing unit tests, reviewed database migrations, and a decision log that traces every design choice. Within AURIVA, it provides the platform's dialogue teaching and speech assessment.

7.5 Limitations and Threats to Validity

The limitations are grouped by the four questions that determine whether research results can be trusted.

Measurement: did the study measure the right thing?

- The 1,500 ms echolalia threshold is a design choice, not a value from the literature, and the Phase 2 prompt is a request with a fixed answer, which can itself increase immediate echoing [13].

- The human pronunciation rating (0-2) and the system score (0-3) measure related but different things, and are compared only after both are reduced to "accepted" or "needs more practice".
- The taxonomy omits the /f/ to /p/ and /z/ to /s/ substitutions reported for Sinhala speakers [12].
- The CMU dictionary gives American English pronunciations, while transcription uses British English and the children speak Sri Lankan English, so a pronunciation acceptable in Sri Lankan English may be marked as an error [24].
- The synthetic-trained model knows only six phoneme error classes; any other class reaching it, such as TH-stopping or cluster reduction, is encoded as all zeros, so that information is lost until the model is retrained on real data.
- The "fast" and "typical" labels are defined by the number of sessions to mastery, which also depends on how often a child attends.
- The observer may have seen the app's feedback before rating some attempts.

Fair comparison: were the conditions controlled?

- During the field study, the synthetic-trained model's influence was switched on, so the order in which words were offered may differ from what the final model would choose. Training labels come from the mastery rules, not from this model, so they are not contaminated.
- The pilot comparison in Table 6.2 used labels later found to be faulty for the struggling class, and its Tier 1 scores used a latency ceiling set from the data rather than the deployed value; the final comparison (Table 6.3) scores all three models on the same rows with their deployed settings.
- Early pilot sessions lack phoneme values and some lack timestamps, so they are not comparable with later sessions and are excluded from training.
- The real-trained model was deliberately constrained (maximum depth 4), whereas the synthetic-trained model was not, so the comparison is between the two models as built, not between the two kinds of data alone.

Generalisation: will it work beyond this context?

- Only three schools took part, each with a small number of autistic children, so the results are formative and not generalisable.
- The design and taxonomy target Sinhala-speaking children; Tamil-speaking children would need their own transfer taxonomy.
- All sessions were supervised by a teacher and the researcher, and testing used two Android-based tablets.
- Non-verbal rows were kept in training although the model never predicts on them; because 17 of the 26 struggling rows are non-verbal, the model partly learned an association between the non-verbal pathway, Abilities words and struggling.

Repeatability: could the results be reproduced?

- An earlier five-fold comparison of seven algorithms was run with a script that is not in the repositories, so only the notebook's comparison, whose outputs are saved, is reported.
- With few children, leave-one-child-out results vary considerably between folds; no significance tests are applied, and every figure is reported with its sample size.
- Speech assessment depends on a commercial service whose models may change over time and were not trained on autistic or Sinhala-accented child speech.
- In mitigation, library versions are pinned, random seeds are fixed, the model file records its library version, and every design decision is recorded in the decision log (Section 3.7).

Implementation limitations. Part of Rule 4, which would reduce Phase 1 exposures after very fast mastery, cannot be reached; the daily wait-time escalation does not apply to Abilities words; a sentence-to-picture matching step designed for Level 2 is not in the deployed flow; the single-voice Level 2 audio was deployed part-way through the field study, so earlier sessions used the two-voice audio; and the component needs an internet connection for speech assessment and personalised audio.

7.6 Implications

For teachers. The reports are evidence to weigh, not verdicts: the trajectory label is a hint, and the explanation shows what it is based on. A flagged transfer error is something to practice, not a failure, and an echoed response is treated as part of learning, not as misbehavior.

For researchers. Learning-analytics models for small special education populations should be trained and evaluated on real data, child by child, should report balanced metrics, and should have their labels audited before any result is trusted. Synthetic data is useful for testing a pipeline and can give a reasonable starting model, but accuracy on synthetic data says little about real performance; a model's value has to be shown on real children.

For developers of educational technology for autistic children. Every model should have a deterministic fallback and a switch that turns it off, no model should be able to change a child's recorded progress, and local language patterns should be designed in rather than treated as errors.

For schools and policy. A curriculum-aligned tool that works for children who do not speak, and that shows teachers its evidence, can support inclusive English teaching where speech and language therapists are scarce.

8 CONCLUSION AND FUTURE DIRECTIONS

8.1 Conclusion

This research set out to help autistic Sri Lankan primary school children learn to use everyday English words and short sentences, not only to recognise or repeat them, and to give their teachers evidence they can trust. The result is the Adaptive Dialogue Learning and Progression Tracker, one of the four components of AURIVA, deployed in the integrated application and used in a supervised field study at three schools.

Level 1 teaches 30 curriculum words through familiarisation, production and contextual awareness, making knowledge of when to use a word an explicit, assessed step. Level 2 teaches the Grade 3 speaking topics through personalised sentences traced to the national activity book. A non-verbal pathway runs through both levels. Three research contributions address the problems the original design could not solve: speech is scored by sound rather than spelling, with Sinhala first-language transfer errors recognised and given partial credit (RC1); probable immediate echolalia is detected from response latency and neither counted as learning nor penalised (RC3); and each child's learning pace is predicted by a two-tier model and explained to the teacher (RC2).

The most important findings are methodological. Accuracy on synthetic data overstated real performance; a small-sample calibration made the model predict the same class for every input; and the early "struggling" labels came from counting faults rather than from children's behaviour. Each failure was found by auditing, not by headline accuracy. Once trustworthy real labels were available, a Random Forest predicted learning pace clearly better than the literature-grounded formula on the verbal attempts it handles (balanced accuracy of about 0.70 against 0.50), although a model trained on real data did not yet improve on the synthetic-trained one. The model sits behind an interpretable formula, a confidence gate and a switch, so it can never change a child's recorded progress.

8.2 Achievement of Objectives

Table 8.1 summarises how far each objective set out in Section 1.5 was achieved.

Table 8.1: Achievement of objectives

Objective	Status	Evidence
O1: Three-phase Level 1 flow with mastery rules and a non-verbal pathway	Achieved	Implemented and deployed (Sections 4.4.1-4.4.3); verified against acceptance criteria and unit tests (Section 5.4)
O2: Curriculum-traced Level 2 with element-based partial credit and the contrastive word-order phase	Achieved	Three topics traced to the activity book (Table 4.7); element scoring and mastery rules (Section 4.4.8)
O3 (RC1): Transfer-aware speech assessment	Implemented; evaluation pending	Phoneme scoring with the Sinhala taxonomy (Section 4.4.4); agreement with the observer pending (Section 6.1.6)
O4 (RC3): Latency-based echolalia handling	Implemented; evaluation pending	Detection and response rules (Section 4.4.5); precision and recall pending (Section 6.1.7)
O5 (RC2): Two-tier trajectory model trained on real data, with explanations	Achieved, with limitations	Real-trained model evaluated leave-one-child-out on 87 rows from 26 accounts, and both learned models beat Tier 1 on the verbal rows (Section 6.1.5); the model is not yet deployed, and the training data includes test accounts
O6: Evaluation through verification, computational evaluation and a field study	In progress	114 passing tests; offline evaluation completed; field study running at three schools since 7 October 2026

8.3 Individual Research and Product Contributions

The researcher was solely responsible for the Adaptive Dialogue Learning and Progression Tracker within AURIVA. On the research side, this covers the three-phase contextual word-learning design, the Sinhala transfer-aware speech assessment method (RC1), the latency-based echolalia rule (RC3), the two-tier explainable trajectory model and its evaluation (RC2), and the audits that exposed the calibration and labelling faults. On the product side, it covers the Level 1 and Level 2 screens and flows, the component's backend services, database tables and migrations, the Python microservice for phoneme scoring, prediction and explanations, the synthetic data generator and the training and evaluation pipeline, the teacher reports, and the personalised Level 2 audio. For the evaluation, the researcher designed the observer sheet, the teacher task usability survey and the feedback questionnaire, and ran the field study.

The shared platform (login, student profiles, integration of the four components and deployment) was provided by the group, with integration and deployment carried out by the group leader. The other three components, Graph Neural Network-Enhanced Adaptive Concept Learning, Motor-Informed Adaptive Handwriting Learning and Phoneme-Aware Adaptive Pronunciation Support, were developed by the other group members. AI assistance in development is declared in Appendix B.

8.4 Future Directions

1. **Retrain and re-evaluate RC2 as real data grows.** Retrain the trajectory model on each new cohort of consented data, decide on evidence whether a three-class model is possible, re-set the confidence threshold from the new model's vote shares, remove the test accounts once enough children's data exists, deploy the real-trained model after supervised observation, and switch the machine-learning path on by default only then. Revisit the Tier 1 weights and cut-offs, which are still untuned placeholders.
2. **Extend the RC1 taxonomy.** Add the /f/ to /p/ and /z/ to /s/ substitutions, replace the American English pronunciation dictionary with Sri Lankan English reference pronunciations, and use the error class to choose the prompt shown to the child, for example showing the mouth position for "th".
3. **Strengthen RC3.** Calibrate the latency threshold from field data, possibly per child; combine timing with intonation and with repetition of the prompt wording, which the literature uses to identify echoes; and test whether shorter, less demanding prompts reduce immediate echoing.
4. **Support Tamil-speaking children** with a Tamil transfer taxonomy and Tamil instruction audio.
5. **Work offline,** using on-device speech recognition and cached audio, so that schools with unreliable internet can use the component.
6. **Evaluate learning outcomes over time** in a longer study with more children and a comparison condition, measuring vocabulary gain and the transfer of words into everyday classroom use.

7. **Reconsider conversation practice carefully:** once basic words are mastered, add predictable, teacher-approved dialogue scripts built only from mastered words, before any return to freely generated conversation.

REFERENCES

- [1] GBD 2021 Autism Spectrum Collaborators, "The global epidemiology and health burden of the autism spectrum: Findings from the Global Burden of Disease Study 2021," *Lancet Psychiatry*, vol. 12, no. 2, pp. 111-121, 2025, doi: 10.1016/S2215-0366(24)00363-8.
- [2] K. A. Shaw et al., "Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years - Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022," *MMWR Surveill. Summ.*, vol. 74, no. SS-2, pp. 1-22, 2025.
- [3] H. Perera, K. Wijewardena, and R. Aluthwelage, "Screening of 18-24-month-old children for autism in a semi-urban community in Sri Lanka," *J. Trop. Pediatr.*, vol. 55, no. 6, pp. 402-405, 2009.
- [4] "Increase in autism among children: Over 9,000 identified in Sri Lanka," *Newsire.lk*, Apr. 10, 2025. [Online].
- [5] H. U. Hashim, M. M. Yunus, and H. Norman, "English as secondary language learning and autism spectrum disorder: The obstacles in teaching and learning the language," *Arab World English J.*, vol. 12, no. 2, pp. 22-30, 2021, doi: 10.24093/awej/vol12no2.2.
- [6] A. M. Mohamed and T. S. Shaaban, "Promoting multilingual skills in children with autism: Parent and teacher perspectives on teaching ESL," *Sustain. Multilingualism*, no. 24, pp. 106-135, 2024, doi: 10.2478/sm-2024-0005.
- [7] A. L. Urrea, V. Fernández-Torres, I. R. Rodríguez-Ortiz, and D. Saldaña, "The use of technology-assisted intervention in vocabulary learning for children with autism spectrum disorder: A systematic review," *Front. Psychol.*, vol. 15, Art. no. 1370965, 2024, doi: 10.3389/fpsyg.2024.1370965.
- [8] X. Jiang and S. R. Cohen, "Supporting autistic children and adolescents' social communication skills through digital technologies: A systematic literature review," *Res. Autism*, vol. 130, Art. no. 202750, 2026, doi: 10.1016/j.reia.2025.202750.
- [9] L. Parsons, R. Cordier, N. Munro, A. Joosten, and R. Speyer, "A systematic review of pragmatic language interventions for children with autism spectrum disorder," *PLoS ONE*, vol. 12, no. 4, Art. no. e0172242, 2017, doi: 10.1371/journal.pone.0172242.
- [10] V.-A. Tran, D.-S. Le, H. H. Hung, and D.-Q. Nguyen, "Improving the accuracy of speech recognition models for non-native English speakers using bag-of-words and deep neural networks," *Sci. Rev.*, vol. 9, no. 2, pp. 10-14, 2023, doi: 10.32861/sr.91.10.14.

- [11] H. R. Bosker, "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies," *Behav. Res. Methods*, vol. 53, pp. 1945–1953, 2021, doi: 10.3758/s13428-021-01542-4.
- [12] C. Samarajiwa and R. M. Abeysekera, "Some pronunciation difficulties of Sinhalese learners of English as a foreign language," *Lang. Learn.*, vol. 14, no. 1-2, pp. 45-50, 1964, doi: 10.1111/j.1467-1770.1964.tb00773.x.
- [13] E. G. Cohn, K. R. McVilly, M. J. Harrison, and L. N. Stiegler, "Repeating purposefully: Empowering educators with functional communication models of echolalia in autism," *Autism Dev. Lang. Impairments*, vol. 7, pp. 1-16, 2022, doi: 10.1177/23969415221091928.
- [14] T. Schodde, L. Hoffmann, S. Stange, and S. Kopp, "Adapt, explain, engage: A study on how social robots can scaffold second-language learning of children," *ACM Trans. Hum.-Robot Interact.*, vol. 9, no. 1, Art. no. 6, 2019, doi: 10.1145/3366422.
- [15] S. Lai et al., "ASD-iLLM: An intervention large language model for autistic children based on real clinical dialogue intervention dataset," in *Findings Assoc. Comput. Linguistics: EMNLP 2025*, 2025, pp. 8058-8079.
- [16] R. Mishra, K. C. Welch, and D. O. Popa, "Human-mediated large language models for robotic intervention in children with autism spectrum disorders," 2024, *arXiv:2402.00260*.
- [17] H. U. Hashim, M. M. Yunus, and H. Norman, "Autism children and English vocabulary learning: A qualitative inquiry of the challenges they face in their English vocabulary learning journey," *Children*, vol. 9, no. 5, Art. no. 628, 2022, doi: 10.3390/children9050628.
- [18] M. L. Edelstein, K. Sloman, and K. Selver, "Effects of demand complexity on echolalia in students with autism," *Behav. Anal. Pract.*, vol. 14, no. 4, pp. 984-990, 2021, doi: 10.1007/s40617-020-00535-7.
- [19] C. Blackburn, M. Tueres, N. Sandanayake, J. Roberts, and R. Sutherland, "A systematic review of interventions for echolalia in autistic children," *Int. J. Lang. Commun. Disord.*, vol. 58, no. 6, pp. 1977–1993, 2023, doi: 10.1111/1460-6984.12931.
- [20] F. Xie, E. Pascual, and T. Oakley, "Functional echolalia in autism speech: Verbal formulae and repeated prior utterances as communicative and cognitive strategies," *Front. Psychol.*, vol. 14, Art. no. 1010615, 2023, doi: 10.3389/fpsyg.2023.1010615.
- [21] S. Filipova, V. Galevska Jovchevski, and D. Shoster, "The role of delayed echolalia produced by children with autism spectrum disorder," *Int. J. Res.*

Publ. Rev., vol. 4, no. 7, pp. 1890-1895, 2023, doi: 10.55248/gengpi.4.723.49127.

- [22] J. Mathée-Scott, G. Corrigan, E. Lorang, Z. Hesse, J. Johnson, and C. E. Venker, "A novel method for examining autistic children's comprehension of individual words produced within delayed echolalia: A proof-of-concept pilot study," *Front. Psychol.*, vol. 16, Art. no. 1681076, 2025, doi: 10.3389/fpsyg.2025.1681076.
- [23] U. S. Jacob, G. O. Adigwe, J. Pillay, A. Osisanya, and S. O. Olatunbosun, "Prompt response, cues-pause-point therapies and gender on management of echolalia among children with autism spectrum disorder," *J. Intellect. Disabil.-Diagn. Treat.*, vol. 12, no. 3, pp. 141–152, 2024, doi: 10.6000/2292-2598.2024.12.03.4.
- [24] T. Kandiah, "Lankan English schizoglossia," *English World-Wide*, vol. 2, no. 1, pp. 63-81, 1981, doi: 10.1075/eww.2.1.05kan.
- [25] A. Parakrama, *De-Hegemonizing Language Standards: Learning from (Post)Colonial Englishes About "English"*. Basingstoke, U.K.: Macmillan, 1995.
- [26] L. W. Anderson and D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY, USA: Longman, 2001.
- [27] R. Gray and D. M. Muldoon, "Teaching receptive vocabulary to two autistic children: A replicated, clinic-based, single case experimental design," *Autism Dev. Lang. Impairments*, vol. 9, pp. 1-15, 2024, doi: 10.1177/23969415241258699.
- [28] R. Cole et al., "Perceptive animated interfaces: First steps toward a new paradigm for human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1391-1405, Sep. 2003, doi: 10.1109/JPROC.2003.817143.
- [29] M. Milne, M. H. Luerssen, T. W. Lewis, R. E. Leibbrandt, and D. M. W. Powers, "Development of a virtual agent based social tutor for children with autism spectrum disorders," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2010, doi: 10.1109/IJCNN.2010.5596584.
- [30] B. Kim, "Using selective simulation to create digital therapeutics for educating social skills to children with high-functioning autism spectrum disorder," in *Proc. IEEE Games, Entertainment, Media Conf. (GEM)*, 2022, doi: 10.1109/GEM56474.2022.10017448.
- [31] C. Deng et al., "ASD-Chat: An innovative dialogue intervention system for children with autism based on LLM and VB-MAPP," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2025, doi: 10.1109/IJCNN64981.2025.11227794.

- [32] Ashwini B, V. Narayan, and J. Shukla, "SPASHT: Semantic and pragmatic speech features for automatic assessment of autism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, doi: 10.1109/ICASSP49357.2023.10095135.
- [33] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253-278, 1994, doi: 10.1007/BF01099821.
- [34] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, Art. no. e1355, 2020, doi: 10.1002/widm.1355.
- [35] W. Farzana, F. Sarker, T. Chau, and K. A. Mamun, "Technological evolvment in AAC modalities to foster communications of verbally challenged ASD children: A systematic review," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3055195.
- [36] P. V. Kumar, D. Balamurugan, K. Suriya, and R. Reshma, "An improved learning experience for English language using big data analytics and AAC," in *Proc. Int. Conf. Sustain. Comput. Data Commun. (ICSCDS)*, 2023, pp. 347-353, doi: 10.1109/ICSCDS56580.2023.10104901.
- [37] K. G. Ranaweera, J. Praneetha, W. P. S. De Silva, K. R. B. N. Kempitiya, I. Gamage, and T. Thilakarathna, "Personalized and interactive learning platform for students with autism spectrum disorder," in *Proc. 5th Int. Conf. Adv. Comput. (ICAC)*, 2023, pp. 828-833, doi: 10.1109/ICAC60630.2023.10417404.
- [38] H. Khosravi et al., "Explainable artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 3, Art. no. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.
- [39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765-4774.
- [40] R. Aburukba, F. Aloul, A. Mahmoud, K. Kamili, and S. Ajmal, "AutiAid: A learning mobile application for autistic children," in *Proc. IEEE 19th Int. Conf. e-Health Netw., Appl. Serv. (Healthcom)*, 2017, doi: 10.1109/HealthCom.2017.8210770.
- [41] 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 *Proc. 8th Int. Conf. Informat. Comput. (ICIC)*, 2023, doi: 10.1109/ICIC60109.2023.10381949.

- [42] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quart.*, vol. 28, no. 1, pp. 75-105, 2004.
- [43] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45-77, 2007.
- [44] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 23-30.
- [45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.
- [46] J. Sauro and J. S. Dumas, "Comparison of three one-question, post-task usability questionnaires," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2009, pp. 1599-1608.

GLOSSARY

[Optional. Define specialised terms a reader may not know.]

APPENDICES

Appendix A: Research Instruments or Extended Results

[To insert: Session Observer Sheet, Teacher Task Usability Survey, and Teacher Feedback Questionnaire (without its Section B).]

Appendix B: Academic Integrity and AI Use Disclosure

This appendix declares every use of artificial intelligence (AI) tools in this research, what each was used for, and how its output was checked. All design decisions, judgements and claims in this thesis are the researcher's own, and the researcher takes full responsibility for them.

After the features of the component were implemented, the bugs found during testing and auditing were fixed, by an AI coding agent, Claude Code (Anthropic), working from the researcher's task specifications in the three repositories. Each feature and each fix was then reviewed separately by the researcher before work moved on to the next one. The review consisted of:

- checking the change against the acceptance criteria in its specification;
- syntax and module-loading checks, and the automated unit tests (Section 5.4);
- a live test of the change on the tablet;
- for database changes, reading the migration and running it manually; the coding agent never ran migrations.

No change was accepted without this review, and work on the next feature or fix started only after the previous one had passed it.

Teaching media. The Level 1 scenario videos were generated with Google Flow and Grok, and the illustrations with Gemini and ChatGPT (Section 3.2.2). Every asset was reviewed by the researcher for accuracy and cultural suitability and was regenerated if it contained errors.