applied sciences

MDPI

Article

Personalized AI-Directed Tutoring for Oral Proficiency Enhancement in Language Education

Pranav Tushar [1], Bowen Zhang [1,2], Indriyati Atmosukarto [1], Donny Soh [1], Rong Tong [1] and Ian McLoughlin [1,3,*]

[1] ICT Cluster, Singapore Institute of Technology, Singapore 828608, Singapore; indriyati@singaporetech.edu.sg (I.A.); tong.rong@singaporetech.edu.sg (R.T.)
[2] College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore
[3] National Engineering Research Center of Speech and Language Information Processing, The University of Science and Technology of China, Hefei 230052, China
* Correspondence: ian.mcloughlin@singaporetech.edu.sg

Abstract

Generative AI offers transformative potential for scalable, personalized, and dynamic language education, particularly in enhancing oral proficiency among young learners. However, effective deployment remains challenging due to limited resources for some languages, the need for age-appropriate content and tools, and the importance of respecting cultural relevance. In this paper, we introduce LEARN (Language Evaluation via question Answer generation from caRtooNs), a culturally grounded multilingual visual dialogue system designed to support oral proficiency in three of Singapore's official languages: Mandarin, Bahasa Melayu, and Tamil. English, as the lingua franca, is excluded. LEARN integrates a teacher-facing module for curriculum-aligned visual question-answering task creation and a student-facing module for voice-driven adaptive dialogue, optimized for children's speech. Unlike existing platforms, LEARN prioritizes cultural relevance and low-resource language support, helping address gaps in heritage language preservation. Pilot studies with students demonstrate significant improvements in engagement and vocabulary acquisition. Designed for classroom as well as home use, LEARN presents a scalable AI-driven language tutoring framework.

Keywords: language learning; AI tutoring; L2 acquisition; dialogue system; low-resource language learning

1. Introduction

The advent of generative AI (GenAI) has provided new opportunities for scalable, adaptive interaction in domains such as creative generation, productivity enhancement, and education [1–6]. The effectiveness of GenAI is enhanced by ease of use through the ability to interact using natural language. Meanwhile, events such as the COVID-19 pandemic have led to acceptance and growing demand for robust digital learning tools [7,8]. While GenAI may not be able to replace traditional pedagogies or teachers at present, it can serve in a complementary fashion. Interactive dialogue and real-time personalization provide a potential educational experience akin to having an individual private tutor. Emerging support for multilingual interaction means that language learning has become a core application aim for GenAI in education [9].

Singapore, with a multi-lingual population primarily comprising three main ethnicities, as well as those from a myriad of other backgrounds, has recognised the potential of

check for updates

Academic Editor: Carlos Delgado

Received: 3 February 2026
Revised: 22 February 2026
Accepted: 24 February 2026
Published: 28 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

AI-enabled learning in fulfilling national priorities. The Ministry of Education mandates that all students must learn a "mother tongue" ("Mother tongue" is the legal term that refers to one of the historical languages of ethnicity, and the terminology used by the funding agency that supported this work. We adopt the terminology in this manuscript, but acknowledge that it carries potential for gender bias), which involves a separate teaching class in either their own historic language of ethnicity, or a choice from the three main languages: Mandarin Chinese, Malay and Tamil. Mother tongue language (MTL) competency is a requirement at the primary level in Singapore, and forms part of the educational assessment for all students. Meanwhile, English is the dominant language of education, work and interaction in many Singaporean households, where it has largely supplanted the historic languages of ethnicity ("mother tongues"). Children often receive limited exposure to their MTLs outside of their MTL class, resulting in reduced heritage language maintenance [10]. Hence, GenAI-powered tutoring systems to support MTL acquisition and retention, as well as promote linguistic diversity at scale [11], would be of great value.

Despite advancements in speech and language technologies, deploying Intelligent Tutoring Systems (ITSs) in real-world classrooms remains challenging. Current approaches typically lack age-sensitive content controls and struggle to adapt to multilingual or code-switching scenarios [7,12]. LLMs can sometimes generate hallucinated, biased, or culturally inappropriate outputs, which are difficult for educators and caregivers to verify, raising concerns about safety and reliability [13,14]. Meanwhile, ASR models frequently underperform on children's speech [15], especially when dealing with accent variations, under-represented low-resource languages in training data [16,17], and background noise in educational settings. The absence of pedagogical scaffolding and personalized feedback mechanisms also hinders support for diverse learning pathways.

The proposed LEARN system bridges these gaps between low-resource languages, automated dialogue systems for second-language acquisition and children's speech interfaces, while aiming to make language learning engaging and accessible for young learners. By combining AI-driven dialogue and appealing visual storytelling, the system assists Primary 1–2 students to improve MTL skills through a visual question-answering (VQA) pedagogy [18]. LEARN contains two key modules: a teacher module to support curriculum-aligned task design and learner tracking, and a student module for real-time, adaptive interaction. Students interact with the system using voice, while teachers interact via a dashboard. Figure 1 illustrates a student module interaction sequence where the spoken dialogue is represented in text boxes.

For Singaporean families, fluency in MTLs is more than a skill as it provides a bridge to cultural heritage and family traditions, for example, in being able to communicate with grandparents.

LEARN aims to strengthen that bridge using vision, speech, and text modalities to build oral proficiency in Mandarin, Bahasa Melayu, or Tamil. The architecture of LEARN is designed to be highly scalable and extensible, while delivering an interaction that is both fun and meaningful. Initial findings, presented below, reveal promising levels of engagement, where students eagerly participated in dialogue to expand their vocabulary and oral proficiency.

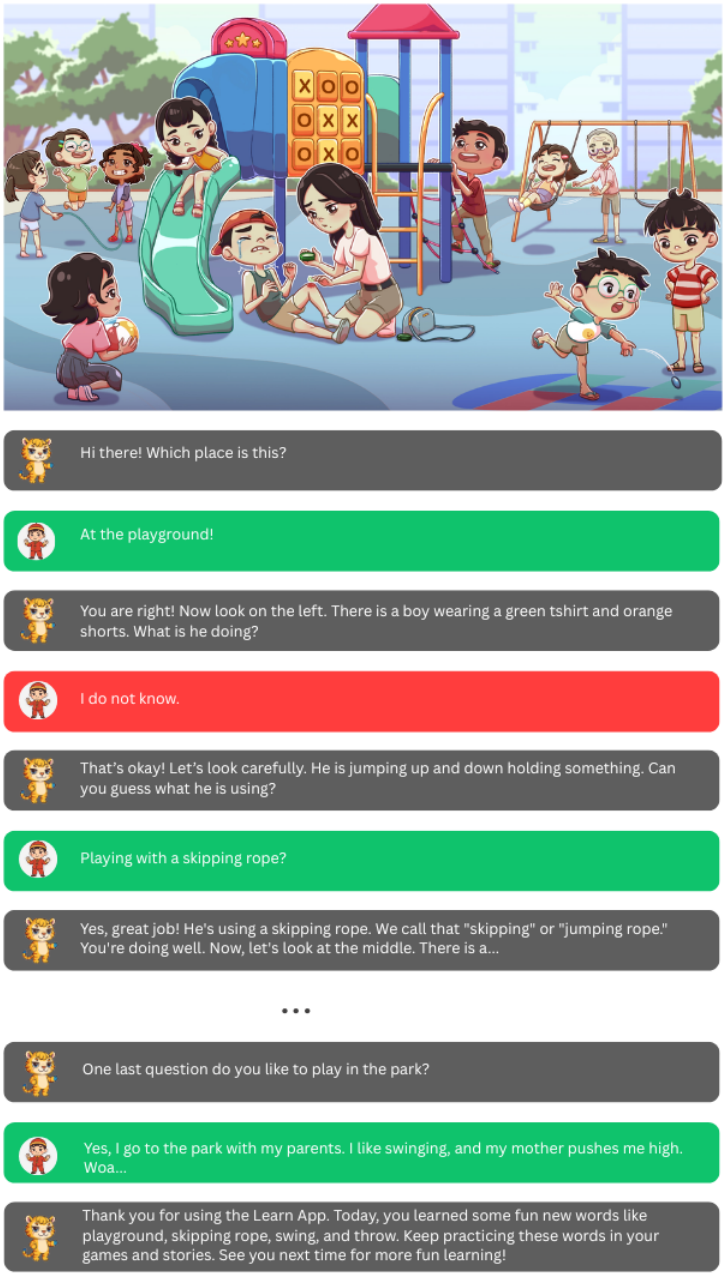


Figure 1. An illustration of the VQA paradigm where LEARN guides a student to enhance vocabulary skills by introducing and practising new words through an adaptive dialogue. For illustrative purposes, the verbal dialogue is presented in text boxes. Correct answers are in green, incorrect ones are in red.

2. Related Works

Recent research in AI-driven education has increasingly focused on integrating generative models into tutoring systems that support adaptive and personalized learning at scale. Moving beyond early rule-based Intelligent Tutoring Systems (ITSs), modern approaches leverage LLMs to dynamically generate instructional content, assess learner responses, and adapt feedback in real time. Advances in prompt engineering, fine-tuning, and retrieval-augmented generation have enabled curriculum-aligned tutoring and assessment across domains such as science [19], programming [20,21], mental health awareness [22], and include automated evaluation through AI teaching assistants [23–25]. Studies suggest that such systems can approximate aspects of one-to-one tutoring by tailoring prompts and explanations to learners' proficiency levels, learning trajectories, and error patterns [26].

Yet, at the same time, the open-ended nature of LLM generation raises concerns regarding pedagogical consistency, factual reliability, and suitability for young learners. This has motivated a shift towards constrained generation pipelines and human-in-the-loop validation in educational deployments [13,14], a solution shared by LEARN.

Within language education, LLM-powered conversational agents have been explored as tools for both first and second-language practice, providing learners with increased opportunities for interactive dialogue beyond the classroom [27,28]. Commercial platforms such as Duolingo, Stimular, and Speak2Me combine LLMs with speech technologies to offer structured learning pathways and feedback. While these systems demonstrate gains in engagement and fluency, most remain text-centric or are primarily designed for adult learners.

Oral proficiency training for children clearly introduces additional challenges, as it requires accurate speech recognition, age-appropriate and relevant feedback, and the subjects' sustained engagement over short interaction cycles. ASR systems trained predominantly on adult speech often perform poorly on children's speech due to acoustic mismatch, developmental articulation differences, and pronunciation variability [15]. These issues are further amplified in low-resource languages, where limited annotated data and dialectal variation significantly degrade recognition accuracy [16] and are very challenging in terms of automated feedback.

Recent work has shown that domain-specific data collection and fine-tuning can substantially improve ASR performance for child speakers and low-resource languages [17]. However, relatively few studies have integrated such speech technologies into end-to-end tutoring systems evaluated in real-world classrooms. This gap limits understanding of how speech-enabled AI tutors perform under realistic conditions, particularly for language learning contexts. The LEARN system and our analysis of its deployment performance help to start filling that gap.

Parallel to advances in speech and dialogue, multimodal learning has gained prominence as a means of enhancing comprehension and engagement for early learners. Grounding language instruction in visual contexts has been shown to support vocabulary acquisition and reduce cognitive load by anchoring meaning in concrete, perceptually salient cues. VLM and VQA frameworks provide a foundation for generating and evaluating language grounded in images [18,29]. Recent VLMs extend these capabilities by supporting dialogue that jointly reasons over visual and linguistic inputs, enabling image-grounded conversational tutoring in real-world contexts [30,31]. Despite this promising progress, many existing systems rely on generic image datasets and lack cultural grounding, which is important for engagement in diverse educational settings.

A notable trend since 2024 is the emergence of child-focused language tutors that explicitly combine LLM-based dialogue with pedagogical orchestration and age-appropriate interaction design. For example, SingaKids [32] implements picture-description pedagogy for early primary learners by integrating dense image captioning, multilingual ASR, dialogic interaction, and adaptive scaffolding across English, Mandarin, Malay, and Tamil. While SingaKids demonstrates the feasibility of multimodal, multilingual tutoring for children, it also highlights persistent challenges, including inconsistent performance across languages, reliance on generic visual representations, and limited mechanisms for curriculum-specific authoring and teacher oversight. In a related direction, Xiao et al. [33] report an LLM-enhanced bilingual conversational agent embedded within an interactive e-book to support children's dialogic reading, where adaptive prompting and feedback are used to encourage oral participation. Their findings underscore that child-facing language systems must be designed not only for linguistic performance, but also for parental and teacher perceptions of educational value, usability, and safety.

Taken together, prior work demonstrates the growing promise of LLM-driven tutoring for supporting interactive oral language practice, while also revealing a central design tension: open-ended generation must be carefully balanced with structured pedagogical constraints, cultural relevance, age-appropriate scaffolding, and safeguards against misleading or inappropriate outputs. In settings like Singapore, where MTL learning depends on contextual engagement, these gaps highlight the need for culturally relevant and age-appropriate AI-enabled solutions such as LEARN.

3. LEARN System Design

Contemporary theories of second-language acquisition (SLA) converge on several core principles relevant to system-guided learning. Krashen's Input Hypothesis [34] proposes that language development occurs when learners receive comprehensible input that is slightly beyond their current proficiency, formalised as the *i+1* principle. Vygotsky's Zone of Proximal Development (ZPD) [35] further emphasises that learners progress most effectively when supported through tasks just above their independent capability. Scaffolding temporary adaptive instructional support enables students to bridge these gaps through hints, prompts, or reformulations, with assistance gradually fading as autonomy increases.

Usage-based approaches [36] complement these views by highlighting that acquisition is driven by repeated exposure to meaningful linguistic usage. Learners internalise grammatical patterns and semantic associations through frequent, reliable encounters with constructions. Theories additionally stress the importance of interaction and multimodality. Long's Interaction Hypothesis [37] argues that comprehensible input is most beneficial when embedded in exchanges that prompt learners to negotiate meaning through clarification requests, expansions, or self-correction. Research on dual coding [38] demonstrates that pairing verbal information with imagery produces complementary visual and linguistic memory traces that enhance retention and retrieval.

LEARN operationalises these SLA principles by scaffolding tasks within each student's ZPD, maintaining *i+1* comprehensibility through adaptive questioning, rephrasing, and vocabulary recaps. Its Dialogue Manager fosters active production by dynamically adjusting follow-up prompts to student responses, encouraging negotiation of meaning. Finally, the integration of culturally grounded cartoon imagery with verbal prompts leverages dual coding to reinforce comprehension and memory.

Cartoons were used rather than photographs to reduce unnecessary details, such as complex backgrounds, uneven lighting variation and confusing textures. In addition to reducing visual clutter, cartoons enabled scenes to be simplified and constructed to direct attention to the key learning elements, actions and spatial distribution. In this, we were guided by prior work [39] on the pedagogical value of cartoons. We also considered the visual appeal of the cartoons to be valuable in maintaining student engagement, which we assessed experimentally below.

The following sections further detail LEARN's system architecture: first, the teacher-facing question generation module in Section 4, and then the student-facing dialogue-based learning environment in Section 5.

4. LEARN Architecture: Teacher Module

The teacher module allows educators to create and manage oral tasks. Given a cartoon image, it utilises a VLM to generate culturally relevant, age-appropriate VQAs in the selected MTL.

To identify a suitable model for this pipeline, previous work [40] conducted a comparative evaluation of several open source VLMs on culturally grounded reasoning tasks situated in Singaporean contexts. Qwen2.5-VL-7B-Instruct [41] emerged as the preferred

model, offering the strongest performance below 10 billion parameters. Although some closed-source models demonstrated higher cultural sensitivity, Qwen2.5-VL-7B-Instruct provided a practical trade-off between accuracy, computational efficiency, and reproducibility. Its relatively compact architecture enabled complete pipeline deployment on a single NVIDIA A6000 GPU, ensuring scalable and cost-effective inference suitable for real-time educational applications. Figure 2 illustrates the overall architecture of the teacher module.

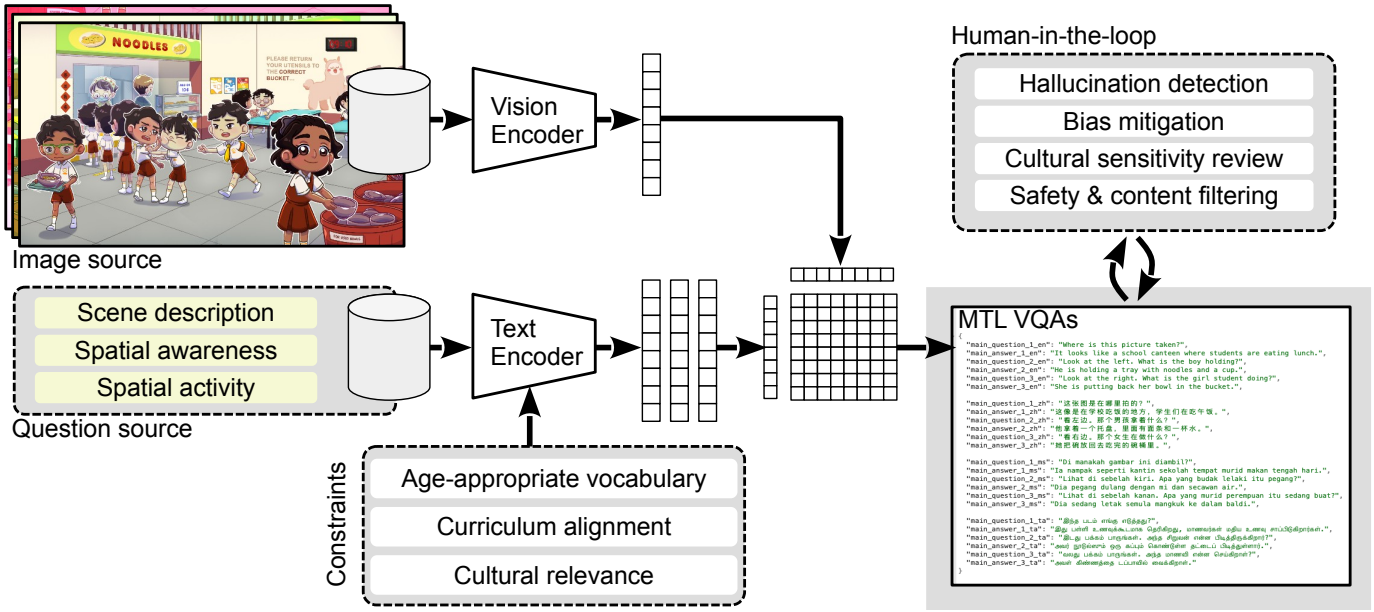


Figure 2. Overview of the teacher module in LEARN. Educators upload images (top left) for automated VQA generation via a VLM that is constrained in terms of curriculum alignment, age-appropriateness, and cultural relevance. Human–AI collaboration is used for an assessment of safety, bias, hallucination and cultural sensitivity before the VQA list is generated in each MTL.

4.1. Image Activity Detection and Grounding

Task creation begins with educators choosing cartoon images around which the VQA sessions will be scaffolded. Student familiarity with the subject matter is important (so that vocabulary and language skills are tested, rather than recognition) [39]. Hence, the cartoons in LEARN depict everyday human activities in common local Singaporean environments such as MRT (train) stations, hawker centres (multi-vendor food courts), playgrounds and schools. Although generative AI content creation tools have demonstrated impressive abilities in recent years, the LEARN team employed a full-time artist. The artist ensured the precise creation of culturally relevant images that support the required pedagogy, while balancing representation in multiracial Singapore. In total, the artist created a dataset of 2500 target images from which educators can select.

The selected images were passed through a structured multi-stage visual analysis pipeline composed of three key components: human activity detection, spatial segmentation, and spatially grounded activity description [42]. This pipeline also produced a structured scene representation by dividing the visual input into three spatial zones: left, middle, and right. This structured scene is the basis for VQA pair generation.

4.2. Curriculum-Aligned VQA Generation

To support the generation of age-appropriate VQA tasks, we first impose a spatial segmentation structure requirement on scene description inputs to the VLM, and then use a curated few-shot prompting strategy based around an educator-provided vocabulary. This yields five structured questions per image:

- A question capturing the overall scene;
- Three questions grounded in specific spatial regions of the image;
- One question aiming to elicit a personal response from participants.

These questions collectively form the core instructional content.

Curricular alignment is achieved by an educator-supplied word bank containing target vocabulary, alongside the curated dataset of cartoon images discussed above. Several strategies were evaluated for question generation, including zero-shot [43], few-shot [44] and LoRA fine-tuning [45] of the VLM. Experienced educators reviewed the results of each. Few-shot prompting was found to best support the curriculum-specific vocabulary lists and achieve the most reliable outcomes. To further enhance language accessibility for young learners, we simplified the vocabulary by replacing less familiar terms with more interpretable alternatives. For example, *"vacuum cleaner"* would be reformulated as *"a machine to clean the floor"*. This was accomplished with LLM-guided detection and prompting, with human decision making. In addition, negative prompting techniques were used to minimize complex language and ensure that generated content was both age-appropriate and aligned with instructional goals.

4.3. Multilingual VQA

Initial development of the multilingual pipeline followed a two-step approach where VQA pairs were first generated in English and then translated into the target MTL using an instruction-tuned LLaMA model [46]. While this method was very flexible and enabled rapid scalability to other languages with good consistency, it tended to introduce cascading translation errors. These affected fluency, and the eventual performance is highly dependent upon the performance of the translation engine. The translation engine is required not only to convert between languages but also maintain cultural relevance and safety, which is difficult in practice. Furthermore, the extended pipeline containing a translation module could be slow, introducing delays that affected the real-time streaming interaction and delayed TTS rendering. A simple illustration of this is to consider sentence grammar where the final word in a source language sentence need to be translated before the initial word in the target language can start to be rendered, e.g., "The whole family, their dog, and their friends are playing in the park today" and "今天，全家人、他们的狗和朋友们都在公园里玩耍。"

To improve output quality and speed, a direct multilingual VQA generation process was developed using language-specific prompts, i.e., the entire pipeline operated in the target language. This worked well in practice for Mandarin and Bahasa Melayu, but not for Tamil.

4.4. Low-Resource Tamil VQA

Direct Tamil VQA generation was unreliable due to frequent hallucinations and semantic errors, reflecting the comparative state of maturity of low-resource versus high-resource language tools, and the urgent need for research in low-resource language technology.

To ensure that VQA material is of good quality, in practice, sentences were created in English and then KrutrimTranslator [47] was used, which achieved a four times latency reduction over IndicTrans2 [48], slightly alleviating pipeline delays caused by incorporating a translation step. The observation of cascading translation errors in Section 4.3 above remains, but these were judged to be less common than errors in direct Tamil VQA generation. In all cases, the consequence of VQA errors is additional checking required by the educators.

The authors note that, as language models continue to improve in fidelity and reliability, especially for low-resource languages, it is likely that direct creation in Tamil will become viable in the short- to medium-term.

4.5. Human AI Collaborative Feedback

To ensure pedagogical and cultural integrity, we implemented a verification framework that evaluated VQA outputs for age-appropriateness and the absence of harmful or biased content. Originally designed for manual educator review, deployment constraints led to a hybrid human–AI validation process using the VLM. Depending on the configuration, each VQA pair is reviewed by either human experts or the model. Validated outputs were serialized into structured JSON, supporting both educator-in-the-loop scenarios and scalable, low-supervision deployment.

5. LEARN Architecture: Student Module

The student module provides a voice-driven dialogue experience that dynamically adapts to learner responses in real time. Each session begins with a TTS-generated main question, followed by a student's spoken response in their MTL, transcribed using ASR. The Personalized Dialogue Manager (PDM), powered by Qwen2.5-VL-7B-Instruct, classifies response intent and determines the next action, whether to give feedback, scaffold with follow-up prompts, or proceed. The aim is to ensure linguistic accuracy, contextual relevance, and a personalized learning flow. Figure 3 presents the overall architecture of the student module.

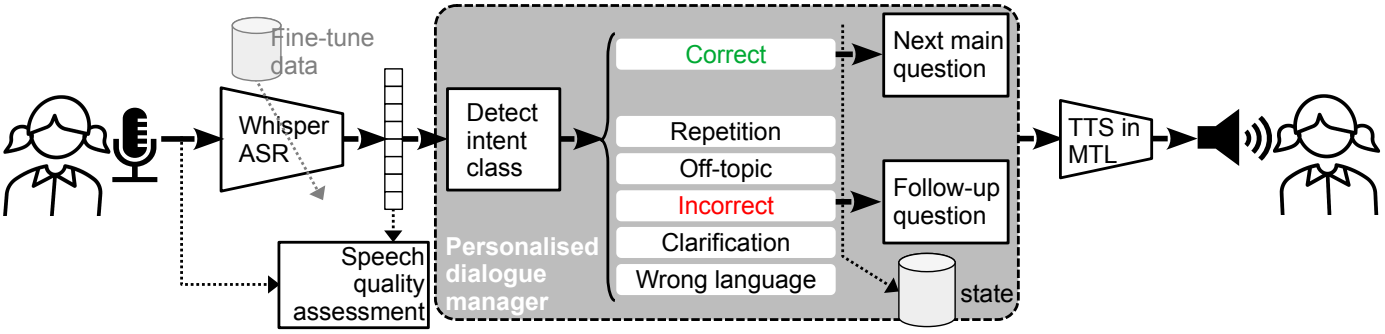


Figure 3. Architecture of the student module. The system delivers a voice-driven adaptive dialogue experience powered by ASR, TTS, and PDM.

5.1. Children ASR

Student responses in their selected MTL are captured as spoken input and transcribed through an ASR module for downstream dialogue processing. In early deployments, the Whisper-Large-V3 model [49] was used. However, it exhibited limited robustness to the acoustic variability of children's speech, which is often characterised by age-related articulation differences, higher pitch ranges, and regional accent variation [15]. These challenges were particularly pronounced in lower-resource MTLs such as Tamil and Bahasa Melayu.

Prior work has shown that task-specific data collection and fine-tuning can substantially improve ASR accuracy for low-resource and child-directed speech domains [17]. Following this methodology, we collaborated with 10 local primary schools to curate a speech dataset comprising approximately 30 h of audio from 659 children aged 7 to 9 across Mandarin, Bahasa Melayu, and Tamil. The dataset includes 1318 recording sessions, with gender representation of 44% boys and 56% girls. All recordings were annotated at both phonetic and semantic levels by trained linguists in each respective MTL to ensure high-quality transcriptions.

To ensure a fair evaluation of generalisation to unseen speakers, all ASR experiments employed a per-speaker train-test split. No child's recordings appeared in both the training and evaluation sets, preventing speaker leakage and avoiding overly optimistic performance estimates. The Whisper-Medium-V3 model [49] was then fine-tuned using LoRA with an 80–20 training–evaluation split defined at the speaker level. The dataset collected for training and fine-tuning is described in Zhang et al., 2025 [50].

This fine-tuning procedure produced consistent reductions in both Character Error Rate (CER) and Word Error Rate (WER) across all three MTLs, as shown in Table 1. Marked improvements in recognising children's speech were achieved for each language, although the Tamil ASR performance was still much poorer than the other languages.

Table 1. Comparison of baseline and fine-tuned ASR model performance on speech data across MTL. Lower numbers indicate better performance.

Language	Baseline CER/WER ↓	Finetuned CER/WER ↓
Mandarin	29.4%	7.5%
Bahasa Melayu	63.6%	10.1%
Tamil	79.5%	35.0%

5.2. Personalized Dialogue Manager (PDM)

The PDM delivers an adaptive learning experience through a mobile or web interface in which each question is vocalized using an MTL TTS engine. The main questions, curated by the teacher module, are accompanied by one to three personalized follow-up questions as noted above, and are designed to maintain student engagement to reinforce learning. To enhance vocabulary development, the system incorporates an intent classification module that assigns each student response to one of six distinct categories: correct, incorrect, clarification request, off topic, repeated input, and language mismatch.

For *correct* responses, PDM acknowledges the input and proceeds to the next question. *Incorrect* answers trigger up to two follow-up questions designed to scaffold understanding (e.g., "*Look at the girl's hand, what is she holding?*"), with encouraging feedback throughout (e.g., "*Great try! Let's look again.*"). *Clarification* requests prompt additional context before reposing the question, while *off topic* replies are acknowledged and gently redirected. To preserve dialogue integrity, PDM flags *repetition* (verbatim echoes) of the prompt, something that could lead to hallucinated validations if missed. The *wrong language* class flags responses in languages other than the expected MTL. Although ASR models are fine-tuned for specific languages, cross-lingual transcription errors (e.g., English detected in a Bahasa Melayu session) can often occur in such a multilingual context. Unlike general-purpose dialogue systems, LEARN is focused on MTL oral proficiency; thus, responses in the wrong language are not accepted, even if semantically correct. Instead, the PDM scaffolds the student back to the intended language, reinforcing consistent language use. Each session concludes with a vocabulary recap (e.g., "*Today, you learned words like 'cleaning,' 'plane,' and 'eating'.*") to consolidate learning.

TTS

All dialogue responses are converted into audio using language-specific TTS engines to support voice-based interaction. The pipeline employs MelayuSpeech [51] for Bahasa Melayu, PaddleSpeech [52] for Mandarin, and Indic-TTS [53] for Tamil. Trained on adult speech, these engines serve as a virtual "teacher voice", with clear and consistent articulation.

5.3. Student Response Evaluation

We explored three strategies to evaluate student responses across modalities. The first method, speech-to-speech semantic evaluation, uses ECAPA-TDNN embeddings [54] to compute cosine similarity between student and reference audio. The second, ASR-based evaluation assesses accuracy using WER and BLEU [55]. The third, VLM-based evaluation, currently deployed, uses Qwen2.5-VL-7B-Instruct to assess responses, offering goal-aligned, grounded evaluation beyond text similarity. The VLM approach was chosen for its robustness, with prior work demonstrating its effectiveness in VQA tasks [29], positioning LLMs as reliable evaluators.

6. System Level Evaluation

Before conducting the classroom pilot, a series of system-level evaluations was carried out to validate the quality, reliability, and pedagogical suitability of the key components of LEARN. These internal assessments examined survey design considerations, dialogue accuracy, VQA generation quality, and teacher workload reduction. The results informed iterative refinement of the system prior to deployment in schools.

6.1. Personalized Dialogue Manager Accuracy

To assess the reliability of PDM, we evaluated both intent classification and the appropriateness of follow-up responses. A controlled dataset comprising 60 synthetic student utterances (six intent categories) and 50 follow-up responses was constructed across Mandarin, Bahasa Melayu, and Tamil. Each output was rated on a five-point scale by both language experts and an external large language model (LLM) evaluator.

Table 2 summarises the average scores across the three languages for utterance-level accuracy and follow-up response quality. It is interesting to note that the LLM-as-a-judge correlated very well with the expert.

Table 2. Evaluation of the Personalized Dialogue Manager using 60 synthetic student utterances (six intent categories) and 50 follow-up responses across three mother tongue languages. Each output was rated on a 5-point scale by language experts as well as an external LLM judge (GPT-4O). Scores reflect the accuracy of intent handling and the appropriateness of generated follow-up responses.

Language	Expert (Original Utterances)	LLM-as-a-Judge (Original Utterances)	Expert (Follow-Up Responses)	LLM-as-a-Judge (Follow-Up Responses)
Bahasa Melayu	4.7	4.6	4.7	4.6
Mandarin	4.5	4.4	4.8	5.0
Tamil	3.8	4.2	4.2	4.4

These evaluations guided iterative revisions to the prompt templates, response scaffolding, and feedback logic, ensuring age-appropriate, consistent, and pedagogically reliable system behaviour across languages.

6.2. Quality of Generated VQA Questions

The teacher-facing VQA generation module was assessed through an educator survey involving five language specialists. Participants rated the age-appropriateness and curriculum relevance of automatically generated questions on a five-point Likert scale. The system achieved an average rating of 4.4/5, with educators reporting that most questions required only minor edits. These results indicate that LEARN is capable of producing classroom-ready VQA items with minimal post-editing effort.

6.3. VQA Generation Workflow and Teacher Effort

To quantify teacher workload reduction, we conducted a controlled comparison of manual versus LEARN-assisted VQA creation. Five educators (two Mandarin, two Tamil, and one Bahasa Melayu) created identical sets of five questions with follow-up prompts under both conditions.

Manual creation required between 22 and 65 min (median 30 min), whereas LEARN-assisted authoring required only 3 to 6 min. This corresponds to an estimated 80–90% reduction in preparation time. Educators reported making few corrections to generated content (mean correction effort approximately 3/5), confirming that most outputs were pedagogically suitable with minimal revision.

6.4. Safety Considerations

To ensure safe interactions with young learners, all testing and system evaluations were conducted with a human in the loop. During both development and pilot phases, educators or researchers monitored system outputs in real time to identify any inappropriate, harmful, or misleading content. In addition, a reporting mechanism was established for teachers to flag any safety concerns or unexpected system behaviours. These safeguards ensured that the system remained pedagogically appropriate and aligned with classroom safety expectations throughout the evaluation process.

7. Pilot Deployment and Feedback

The pilot deployment of LEARN was conducted over a two-week period using a six-core NVIDIA A6000 GPU server. The infrastructure supported simultaneous interactions with students across all three language streams and aimed to demonstrate feasibility for classroom-scale deployment.

7.1. Subjective Performance

To evaluate LEARN's real-world impact, the pilot study targeted three key dimensions: student engagement, vocabulary acquisition, and system usability for primary education. In total, 206 Primary 1 and 2 students had parental consent to participate under educator supervision. We conducted post-engagement surveys (Figures 4 and 5) to assess usability and learner experience, while vocabulary acquisition was reported on by the educators. Results revealed strong engagement, especially with image-based tasks, and a positive reception to auditory feedback. Most students reported that the questions were manageable, although a small subset indicated some difficulty with vocabulary. Enjoyment scores were consistently high across all languages and grade levels.

The open-ended feedback given by educators consistently reported strong student engagement, while some cited reduced workload due to the automated question generation and monitoring. Educators noted student vocabulary acquisition, including being better able to produce target words post-test, and this was consistent with the in-session tracking of learned vocabulary and in-session vocabulary recaps. However, due to the relatively short duration of sessions in the pilot deployment, this was not evaluated via a pre/post session comparison.

Each assignment comprised five questions and follow-up responses, typically completed within 5–15 min making the system suitable for both classroom use as well as for at-home learning. Open-ended feedback from parents revealed that they valued the web interface for tracking learning progress and vocabulary growth. Also, the personalization mechanism scaffolds vocabulary learning to suit the progress of each child. For example, unfamiliar words like "climbing" were taught through follow-up prompts such as "repeat after me".

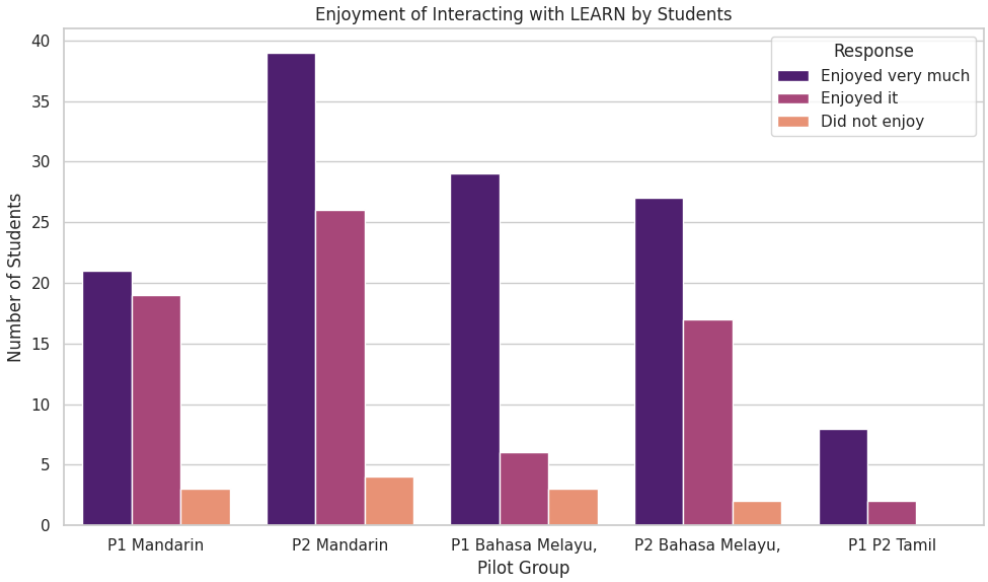


Figure 4. Post engagement enjoyment ratings from 206 Primary 1–2 students across three MTL streams.

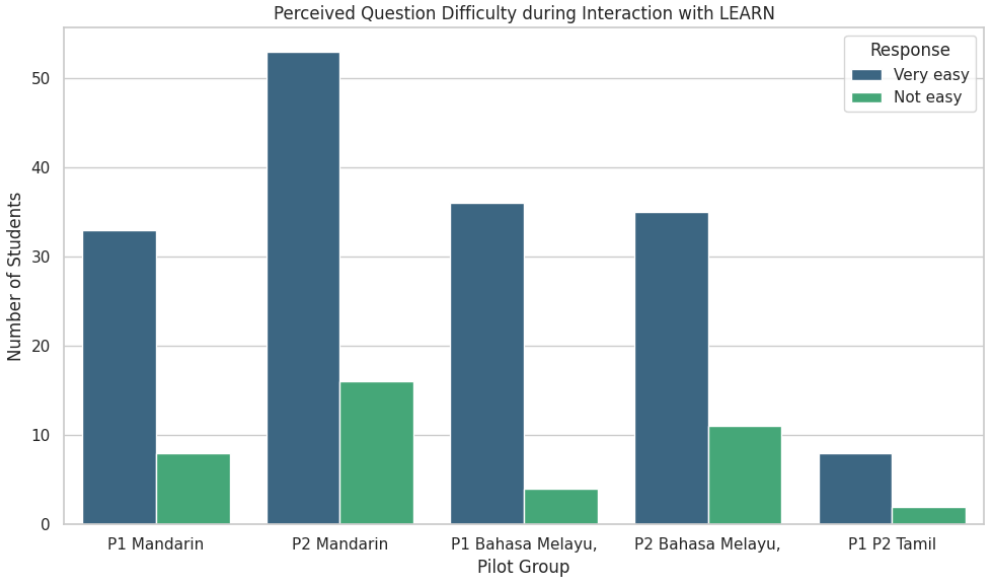


Figure 5. Perceived difficulty ratings reported by 206 Primary 1–2 students across three MTL streams.

One important finding was the system's robustness to off-task responses: when students gave unrelated answers (e.g., *"I want to see my mother"*), the PDM acknowledged them and redirected the conversation. Persistent distraction, however, still required adult intervention.

Another issue noted was that some students defaulted to English due to a lack of vocabulary knowledge. When they were marked incorrect for "language mismatch", but had provided a factually correct answer, frustration was occasionally expressed. This highlights a need for gentle reinforcement of MTL learning expectations.

7.2. Technical Performance

The operational system was deployed via containerized microservices using a primary Triton Inference Server [56] (v.3.5.1) with a fastAPI fallback [57] (v.0.122.1), orchestrated using Kubernetes (v.1.34.2), and exposed through Docker to ensure real-time responsiveness as well as scalability, i.e., compatibility for cloud-based deployment in the future.

The observed performance during the pilot deployment is presented in Table 3, with statistics for each model in the student module reported separately. Performance is shown

for the target 20–30 concurrent users, as well as stress-tested with 100 concurrent users. No model or query failures were experienced during the deployment, but under high load, the system utilised nearly all of the 48 GB available VRAM in the A6000. This is because memory swapping rather than dropping of queries was occurring, which is also consistent with the latency increases (i.e., 4× increase in users leads to 10× the latency). Exploring further, the fine-tuned ASR service had a latency of 0.75 s mean/0.87 s P95 under low load (10 concurrent users), achieving a throughput of 5.32 requests per second. High load latency (50 concurrent users sustained for 300 s) led to 4.97 s mean/6.2 s P95 with a throughput of 7.64 requests per second. Stress testing with 100 concurrent users sustained for 300 s yielded a latency of 9.88 s mean/13.0 s P95 and a throughput of 7.51 requests per second.

VQA generation performance via the teacher module is presented at the bottom of Table 3. This was asynchronous to the student module and could easily be implemented using a separate GPU or cloud service.

Table 3. Pilot deployment statistics on NVIDIA A6000 with AMD EPYC7763. Approximate figures are reported separately for the synchronous student module ASR, TTS and PDM models, plus the teacher module VQA model.

Student Module	**ASR**	**PDM**	**TTS**
Model	Whisper-small, Triton	Qwen2.5-VL 7B text	Section 5.2
20–30 user latency	0.7–1.0 s	0.5–1.5 s	0.3–0.8 s
100 user latency	9–10 s	1–2 s	0.5–1 s
VRAM peak	30 GB	18–20 GB	2.5–11 GB [1]
Failures	0	0	0
Teacher Module	Asynchronous VQA generation (not part of the student pipeline)		
Model	Qwen2.5-VL 7B vision-language		
Performance	VRAM peak ≈ 30–32 GB, latency 3–6 s/image		

[1] Peak VRAM for Malay 10 GB; Mandarin 2.5 GB; Tamil TTS 1 GB & Translate 8 GB.

8. Limitations

While LEARN demonstrates strong potential for multilingual oral learning, several limitations remain. (a) Low-resource languages are poorly served by existing tools. The authors found limitations with Tamil ASR in particular, as well as in Tamil question generation (requiring an intermediate step of English generation followed by translation). This highlights the urgent need for research in under-represented and low-resource languages such as Tamil. (b) ASR systems struggle with code switching. In LEARN, this could lead to misclassifications of multilingual speech. While general ASR in most multilingual settings would need to accept code-switched input, for the purpose of MTL, code switching needs to be detected and flagged. We found it also needs to be flagged appropriately—the answer might be correct, but in the wrong language. This must be handled very differently from an incorrect answer in the right language. It is a disfluency rather than a mistake. (c) While the teacher-in-the-loop overhead was small (i.e., teachers generally did not need to intervene, requiring very little time for monitoring, which was a one-shot process), LLMs and VLMs are known to occasionally generate age-inappropriate or inaccurate responses. AI abilities are inexorably improving, but, for the foreseeable future, systems deployed in sensitive settings such as children's education will still require an element of human filtering.

9. Conclusions and Future Work

This paper has introduced the LEARN, multilingual visual dialogue system for young MTL learners, and evaluated its performance 'in the wild' through a medium-scale pilot

deployment. LEARN is designed to support oral proficiency education for second language learners, the so-called mother tongue languages (MTLs), taught at primary school in Singapore. As a teacher-in-the-loop education system employing LLMs, LEARN constructs pedagogically informed dialogue flows based around attractive and colourful hand-drawn cartoon images, with oversight from teachers and educational professionals. The images and visual question-answering (VQA) dialogues are culturally and pedagogically appropriate and relevant to the daily lives of learners. Survey and engagement results collected during the pilot deployment from 206 primary-aged students demonstrated notable improvements in vocabulary acquisition and oral fluency, as well as high levels of learner engagement.

While this research is situated within the context of Singapore's MTL policy, it is grounded in educational theories and pedagogical approaches that are likely to be easily transferable to other educational settings. More broadly, this work illustrates very clearly the potential for well-designed generative AI-based tools to deliver safe and context-sensitive learning experiences for children's education, in a way that is appreciated by educators and learners alike.

In the future, the authors plan a larger-scale deployment with pre- and post-evaluation of vocabulary. We believe reducing the human-in-the-loop intervention will best be accomplished through a tiered oversight model with: (a) Pre-deployment filtering via teacher module validation in conjunction with automated safety classifiers (i.e., checks for toxicity, age-appropriateness and policy compliance). Vocabulary and structural constraints would also continue to be applied to the outputs. (b) At run-time, automated monitoring of LLM outputs via structured LLM-based validators with possible guardrail-style input/output validation (e.g., for toxicity, PII, grounding checks). When these automated checks flag something, the offending output can be (a) replaced with a safe default (e.g., "Let us try the next question"), or (b) logged for asynchronous review, while allowing the session to continue. Research has shown that an LLM-as-a-judge [58] can effectively complement human risk scoring. Essentially, LLMs perform structured compliance checks to flag sessions for post-session asynchronous review by educators, along with periodic audit sampling. As LLM capabilities increase, higher levels of automation in checking will become viable.

Finally, future research is required in the area of low-resource language tool capability, particularly for ASR and VLM/LLMs. Language tool performance ranking in the LEARN trial correlates with language resource availability. Research is required in improving low-resource language tools. For the Singapore context, this is particularly urgent in the case of the Tamil language.

Author Contributions: Conceptualization, D.S., R.T., I.A. and I.M.; methodology, D.S., R.T., P.T. and B.Z.; software, P.T. and B.Z.; validation, P.T., B.Z., R.T. and D.S.; formal analysis, R.T. and P.T.; investigation, D.S.; resources, D.S. and R.T.; data curation, D.S., P.T., B.Z. and R.T.; writing—original draft preparation, P.T. and I.M.; writing—review and editing, all; visualization, P.T.; supervision, D.S. and R.T.; project administration, D.S.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG award no. AISG2-GC-2022-004).

Institutional Review Board Statement: This research was approved by the Institutional Review Board Statement of Singapore Institute of Technology, with approval number IRB-2023030.

Informed Consent Statement: Informed parental consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used for the research in this article are not available because they contain the identifiable, personal information of child participants.

Acknowledgments: The authors extend sincere appreciation to the subject and language matter experts from the National Institute of Education (NIE), Nanyang Technological University (NTU), and the Singapore Institute of Technology (SIT) for their invaluable advice and support. We also thank the schools that generously hosted and supported the study, the parents and students who participated in the pilot.

Conflicts of Interest: There is no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
CER	Character Error Rate
ECAPA	Emphasized Channel Attention, Propagation, and Aggregation
GPU	Graphics Processing Unit
LEARN	Language Evaluation via question & Answer generation from caRtooNs
LLM	Large Language Model
LoRA	Low Rank Adaptation
MTL	Mother Tongue Language
MRT	Mass Rapid Transit
PDM	Personalized Dialogue Manager
SLA	second-language acquisition
TDNN	Time Delay Neural Network
TTS	Text to Speech
VLM	Vision Language Model
VQA	Visual Question-Answering
WER	Word Error Rate
ZPD	Zone of Proximal Development

References

1. Yan, L.; Greiff, S.; Teuber, Z.; Gašević, D. Promises and challenges of generative artificial intelligence for human learning. *Nat. Hum. Behav.* **2024**, *8*, 1839–1850. [CrossRef] [PubMed]
2. Chakraborty, S. Generative AI in Modern Education Society. *arXiv* **2024**, arXiv:2412.08666. [CrossRef]
3. Hussain, K.; Khan, M.L.; Malik, A. Exploring audience engagement with ChatGPT-related content on YouTube: Implications for content creators and AI tool developers. *Digit. Bus.* **2024**, *4*, 100071. [CrossRef]
4. Antony, V.N.; Huang, C.M. ID. 8: Co-Creating visual stories with Generative AI. *ACM Trans. Interact. Intell. Syst.* **2025**, *14*, 1–29. [CrossRef]
5. Alenezi, M.; Akour, M. AI-Driven Innovations in Software Engineering: A Review of Current Practices and Future Directions. *Appl. Sci.* **2025**, *15*, 1344. [CrossRef]
6. Al Naqbi, H.; Bahroun, Z.; Ahmed, V. Enhancing work productivity through generative artificial intelligence: A comprehensive literature review. *Sustainability* **2024**, *16*, 1166. [CrossRef]
7. Jurenka, I.; Kunesch, M.; McKee, K.R.; Gillick, D.; Zhu, S.; Wiltberger, S.; Phal, S.M.; Hermann, K.; Kasenberg, D.; Bhoopchand, A.; et al. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv* **2024**, arXiv:2407.12687.
8. Azevedo, J.P.W.D.; Rogers, F.H.; Ahlgren, S.E.; Cloutier, M.H.; Chakroun, B.; Chang, G.C.; Mizunoya, S.; Reuge, N.J.; Brossard, M.; Bergmann, J.L. *The State of the Global Education Crisis: A Path to Recovery (Vol. 2): Executive Summary (English)*; World Bank Group: Washington, DC, USA, 2021. Available online: http://documents.worldbank.org/curated/en/184161638768635066 (accessed on 23 February 2026).
9. Xia, Y.; Shin, S.Y.; Kim, J.C. Cross-cultural intelligent language learning system (CILS): Leveraging AI to facilitate language learning strategies in cross-cultural communication. *Appl. Sci.* **2024**, *14*, 5651. [CrossRef]
10. Wenhan, X.; Chin, N.B.; Cavallaro, F. Living in harmony: The negotiation of intergenerational family language policy in Singapore. *Lang. Commun.* **2022**, *82*, 8–27. [CrossRef]

11. Dennison, D.V.; Ahtisham, B.; Chourasia, K.; Arora, N.; Singh, R.; Kizilcec, R.F.; Nambi, A.; Ganu, T.; Vashistha, A. Teacher-AI Collaboration for Curating and Customizing Lesson Plans in Low-Resource Schools. *arXiv* **2025**, arXiv:2507.00456.
12. Jiao, J.; Afroogh, S.; Chen, K.; Murali, A.; Atkinson, D.; Dhurandhar, A. LLMs and Childhood Safety: Identifying Risks and Proposing a Protection Framework for Safe Child-LLM Interaction. *arXiv* **2025**, arXiv:2502.11242. [CrossRef]
13. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
14. Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F.G.; Song, I.; Oh, A.; Augenstein, I. Survey of cultural awareness in language models: Text and beyond. *Comput. Linguist.* **2025**, *51*, 907–1004. [CrossRef]
15. Qian, M.; McLoughlin, I.; Quo, W.; Dai, L. Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM. In *Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*; IEEE: New York, NY, USA, 2016; pp. 1–5.
16. Cibrian, F.L.; Chen, Y.; Anderson, K.; Abrahamsson, C.M.; Motti, V.G. Limitations in speech recognition for young adults with down syndrome. *Univers. Access Inf. Soc.* **2025**, *24*, 2295–2313. [CrossRef]
17. Zhang, B.; Latiff, N.A.A.; Kan, J.; Tong, R.; Soh, D.; Miao, X.; McLoughlin, I. Automated evaluation of children's speech fluency for low-resource languages. In Proceedings of the Interspeech, Rotterdam, The Netherlands, 17–21 August 2025; pp. 1948–1952. [CrossRef]
18. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and Yang: Balancing and Answering Binary Visual Questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
19. Jia, F.; Sun, D.; Looi, C.K. Artificial intelligence in science education (2013–2023): Research trends in ten years. *J. Sci. Educ. Technol.* **2024**, *33*, 94–117. [CrossRef]
20. Dascalescu, S.; Dumitran, A.M.; Vasiluta, M.A. Leveraging Generative AI for Enhancing Automated Assessment in Programming Education Contests. *arXiv* **2025**, arXiv:2506.05990. [CrossRef]
21. Boguslawski, S.; Deer, R.; Dawson, M.G. Programming education and learner motivation in the age of generative AI: Student and educator perspectives. *Inf. Learn. Sci.* **2025**, *126*, 91–109. [CrossRef]
22. Graham, S.; Depp, C.; Lee, E.E.; Nebeker, C.; Tu, X.; Kim, H.C.; Jeste, D.V. Artificial intelligence for mental health and mental illnesses: An overview. *Curr. Psychiatry Rep.* **2019**, *21*, 116. [CrossRef] [PubMed]
23. Tracy, K.; Spantidi, O. Impact of GPT-Driven Teaching Assistants in VR Learning Environments. *IEEE Trans. Learn. Technol.* **2025**, *18*, 192–205. [CrossRef]
24. Denny, P.; MacNeil, S.; Savelka, J.; Porter, L.; Luxton-Reilly, A. Desirable characteristics for ai teaching assistants in programming education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 408–414.
25. Kotsis, K.T. ChatGPT as teacher assistant for physics teaching. *EIKI J. Eff. Teach. Methods* **2024**, 2. [CrossRef]
26. Maity, S.; Deroy, A. Generative ai and its impact on personalized intelligent tutoring systems. *arXiv* **2024**, arXiv:2410.10650. [CrossRef]
27. Kostka, I.; Toncelli, R. Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *Tesl-Ej* **2023**, *27*, n3. [CrossRef]
28. Zhang, Z.; Huang, X. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon* **2024**, *10*, e25370. [CrossRef]
29. Mañas, O.; Krojer, B.; Agrawal, A. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2024; Volume 38, pp. 4171–4179.
30. Castro, G.P.B.; Chiappe, A.; Rodríguez, D.F.B.; Sepulveda, F.G. Harnessing AI for Education 4.0: Drivers of Personalized Learning. *Electron. J. E-Learn.* **2024**, *22*, 1–14. [CrossRef]
31. Bhutoria, A. Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100068. [CrossRef]
32. Liu, Z.; Lin, G.; Tan, H.L.; Zhang, H.; Lu, Y.; Gao, X.; Yin, S.X.; He, S.; Goh, H.H.; Wong, L.H.; et al. SingaKids: A multilingual multimodal dialogic tutor for language learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 1244–1253.
33. Xiao, F.; Li, Z.; Lin, J.; Zou, X.; Yang, D.; Zou, E.W.; Xiong, J. Leveraging an LLM-enhanced bilingual conversational agent for EFL children's dialogic reading: Insights from children, parents, and educators. *Comput. Educ. Artif. Intell.* **2025**, *9*, 100484. [CrossRef]
34. Krashen, S.D. *The Input Hypothesis: Issues and Implications*; Longman: London, UK; New York, NY, USA, 1985.
35. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: Cambridge, MA, USA, 1978; Volume 86.
36. Foster-Cohen, S. Constructing a language: A usage-based theory of language acquisition. *Stud. Second Lang. Acquis.* **2004**, *26*, 491–493. [CrossRef]

37. Long, M. The role of the linguistic environment in second language acquisition. In *Handbook of Second Language Acquisition*; Academic Press: New York, NY, USA, 1996; pp. 413–468.
38. Paivio, A. *Mental Representations: A Dual Coding Approach*; Oxford University Press: Oxford, UK, 1990.
39. Bahrani, T.; Soltani, R. The pedagogical values of cartoons. *Res. Humanit. Soc. Sci.* **2011**, *1*, 19–22.
40. Tushar, P.; Pandey, E.; Austria, L.D.B.; Loo, Y.Y.; Lim, J.H.; Atmosukarto, I.; Lock, D.S.C. MerCulture: A Comprehensive Benchmark to Evaluate Vision-Language Models on Cultural Understanding in Singapore. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*; The Computer Vision Foundation/IEEE Computer Society: Nashville, TN, USA, 2025; pp. 565–574.
41. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2. 5-vl technical report. *arXiv* **2025**, arXiv:2502.13923.
42. Girish, D.; Singh, V.; Ralescu, A. Understanding action recognition in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*; The Computer Vision Foundation/IEEE Computer Society: Piscataway, NJ, USA, 2020; pp. 370–371.
43. Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; Dong, X. Better zero-shot reasoning with role-play prompting. *arXiv* **2023**, arXiv:2308.07702.
44. Ma, H.; Zhang, C.; Bian, Y.; Liu, L.; Zhang, Z.; Zhao, P.; Zhang, S.; Fu, H.; Hu, Q.; Wu, B. Fairness-guided few-shot prompting for large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 43136–43155.
45. Li, Y.; Yu, Y.; Liang, C.; He, P.; Karampatziakis, N.; Chen, W.; Zhao, T. Loftq: Lora-fine-tuning-aware quantization for large language models. *arXiv* **2023**, arXiv:2310.08659.
46. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
47. Srivastava, K.; Zaid, M.; Sowjanya, P.; Adavanne, S. Krutrim Translate. Available online: https://github.com/ola-krutrim/KrutrimTranslate (accessed on 1 December 2025).
48. Gala, J.; Chitale, P.A.; Raghavan, A.K.; Gumma, V.; Doddapaneni, S.; Kumar, A.; Nawale, J.A.; Sujatha, A.; Puduppully, R.; Raghavan, V.; et al. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Trans. Mach. Learn. Res.* **2023**. [CrossRef]
49. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*; ICML: Honolulu, HI, USA, 2023; pp. 28492–28518.
50. Zhang, B.; Latiff, N.A.A.; Tong, R.; Soh, D.; McLoughlin, I. Scsmt: A Multilingual Children's Speech Corpus for Singapore's Mother Tongues. In *Proceedings of the 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*; IEEE: New York, NY, USA, 2025; pp. 543–548.
51. Husein, Z. Malaya-Speech. Available online: https://github.com/mesolitica/malaya-speech (accessed on 1 October 2025).
52. Zhang, H.; Yuan, T.; Chen, J.; Li, X.; Zheng, R.; Huang, Y.; Chen, X.; Gong, E.; Chen, Z.; Hu, X.; et al. PaddleSpeech: An Easy-to-Use All-in-One Speech Toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022.
53. Kumar, G.K.; Praveen, S.; Kumar, P.; Khapra, M.M.; Nandakumar, K. Towards building text-to-speech systems for the next billion users. In *Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2023; pp. 1–5.
54. Desplanques, B.; Thienpondt, J.; Demuynck, K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv* **2020**, arXiv:2005.07143. [CrossRef]
55. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; ACL '02*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
56. NVIDIA Corporation. Triton Inference Server: An Optimized Cloud and Edge Inferencing Solution. Available online: https://github.com/triton-inference-server (accessed on 23 February 2026).
57. Ramírez, S. FastAPI. Available online: https://fastapi.tiangolo.com/ (accessed on 1 October 2025).
58. Xuereb, E. Exploring NLP Models for Generative Storytelling. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.