LLM-based conversational agents for dyslexia treatment: From chatbots to structured tutors

Giulia Valcamonica*
Politecnico di Milano
Milano, Italy
giulia.valcamonica@polimi.it

Giovanni Caleffi*
Politecnico di Milano
Milano, Italy
giovanni.caleffi@polimi.it

Francesco Piferi
Politecnico di Milano
Milano, Italy
francesco.piferi@polimi.it

Marta Tagliani
University of Verona
Verona, Italy
marta.tagliani@univr.it

Maria Vender
University of Verona
Verona, Italy
maria.vender@univr.it

Denis D. Delfitto
University of Verona
Verona, Italy
denis.delfitto@univr.it

Franca Garzotto
Università degli Studi di
Milano-Bicocca
Milano, Italy
Politecnico di Milano
Milano, Italy
franca.garzotto@polimi.it

Abstract

Developmental dyslexia involves persistent difficulties in word-level reading and decoding, requiring sustained linguistic practice that is difficult to maintain without supervision. Although Generative AI offers personalized support, standard Large Language Models (LLMs) often lack the pedagogical and therapeutic knowledge required for linguistic intervention. We present Foxy, a proactive LLM-driven Conversational Agent designed to assist Italian children aged 8–11 with dyslexia during morphological training. Through a modular prompt-orchestration framework and an event-driven architecture, Foxy acts as a specialized tutor that provides scaffolding, limits topic drift, and reduces hallucinations via a verified lexical knowledge base. The system was refined through expert-led co-design and evaluated in a pilot study with 18 educators, and therapists. Results indicate positive perceptions of usability, usefulness, and appropriateness, with recognition of Foxy's motivational benefits. These findings suggest Foxy is a promising tool for dyslexia intervention, pending wider empirical validation.

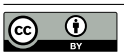
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools*; *Natural language interfaces*; • **Computing methodologies** → Natural language processing.

Keywords

Conversational Agent, Prompt Engineering, LLM, Dyslexia

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CUI '26, Bremen, Germany
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816226

ACM Reference Format:
Giulia Valcamonica, Giovanni Caleffi, Francesco Piferi, Marta Tagliani, Maria Vender, Denis D. Delfitto, and Franca Garzotto. 2026. LLM-based conversational agents for dyslexia treatment: From chatbots to structured tutors. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3816046.3816226

1 Introduction and Background

Developmental dyslexia is a neurodevelopmental disorder characterized by persistent difficulties in word-level reading and decoding, in the absence of intellectual disability, sensory impairment, or inadequate educational exposure [20, 22]. It is widely considered a language-based disorder associated with deficits in phonological processing, which affect the acquisition and automatization of grapheme-phoneme correspondences [8, 29, 31].

Standard interventions for dyslexia primarily involve direct, one-on-one sessions with speech therapists and specialized educators [1]. In clinical and educational settings, professionals support and adapt their explanations to the child's comprehension level [1]. A central component of these interventions is systematic training of the child's phonological skills, and, more recently, of their morphological abilities [3, 17]. Since the core deficit in dyslexia lies in the phonological system (the cognitive mechanism responsible for processing, storing, and retrieving the sound structure of language) [14, 31], exercises targeting phonological awareness aim to strengthen the child's ability to segment and manipulate speech sounds onto their written forms [4, 7]. On the other hand, morphological training (the explicit instruction on the meaning and structure of morphemes) [17] teaches children to recognise the internal structure of words (roots, prefixes, and suffixes) and has proven to be particularly effective for enhancing decoding, especially in people with dyslexia. [12, 24]. Consistent practice with

these exercises is therefore essential, as repeated exposure consolidates the grapheme-phoneme and morpheme-level mappings that underlie fluent reading.

Outside of structured clinical sessions students often rely on compensatory methods to complete their everyday tasks. Text-to-Speech programs, for example, read material aloud to help children overcome reading difficulties [16], but they do not actively train decoding skills. While these tools mitigate daily reading challenges, they are fundamentally passive [25]: they enable task completion but do not reinforce the phonological and morphological competencies that structured therapies attempt to enhance. This creates a gap between the active training children receive in therapy and the passive support available during independent practice.

This gap is a significant challenge because constant language practice is important to treat the underlying deficits [3]. Still, performing morphological and phonological exercises without supervision is challenging. Because of their baseline decoding weakness, these children frequently experience cognitive overload and discouragement, which quickly decreases their motivation [32]. Individuals who practice alone lack the immediate feedback and emotional support provided by a trained tutor [27], and often struggle to fix mistakes and stay engaged in activities without real-time guidance [27]. As a result, there is a major bottleneck: the exercises that might benefit these subjects the most are also the ones they struggle to maintain on their own without expert supervision.

Recent advancements in Natural Language Processing (NLP) and Generative AI (GenAI) have sparked interest in creating technology-based remedies for dyslexia [13, 21]. Historically, AI methods in this sector focused on early screening and detection rather than active intervention [2, 19, 38]. For example, Machine learning methods have been used to identify children at risk of reading difficulties through handwriting analysis, eye-tracking patterns, and speech signal processing [6, 26]. Within dyslexia intervention, AI has been applied in a variety of training settings such as serious games [34], automatic feedback for language learning [37], writing assistance [13] or the creation and generation of linguistic training exercises [33].

Conversational Agents (CAs) have demonstrated potential in education to increase student engagement and sustain interaction over time [10, 36]. However, their application for children with specific language disorders is limited [35]. Large Language Models (LLMs) could close this gap by providing individualized support similar to that of a human specialist [18]. They can generate context-appropriate natural language responses, change explanations dynamically, and preserve multi-turn conversational coherence. These qualities are crucial for engaging children during scheduled activities.

Still, deploying Conversational Agents for children with dyslexia poses considerable obstacles. Out-of-the-box LLMs frequently lack the necessary pedagogical and therapeutic knowledge needed for linguistic intervention. The use of a communication style that is too complicated for children, failure to sustain the child's attention, and limited knowledge of certain morphological exercise mechanics might result in hallucinations or cognitive overload for the user [15].

To address these limitations, we have designed and developed *Foxy*, a proactive LLM-driven assistant built specifically for helping children with dyslexia with Italian morphological exercises. Instead of acting like a generic chatbot, Foxy is embedded in dyslexia-specific interactive exercises and is smoothly integrated with the interaction context. Using a personalized modular prompt engineering method, we constrain the LLM to utilize accessible, child-appropriate, but scientifically accurate language, prevent subject drift, and intervene proactively with emotional and linguistic support when the child struggles.

2 Design of the LLM-Based Morphological Tutor

The intervention method supported by Foxy was designed by dyslexia researchers and specialists, and aimed to strengthen dyslexic children's *morphological competence*, that is, their ability to recognize and manipulate the internal structure of words. In particular, the training focused on helping children identify *base words* (the core lexical element that carries the main meaning, i.e., *dente* "tooth") and *derivational suffixes* (word endings that attach to a base to create a new word with a related meaning, i.e., *-ista* in *dentista* "dentist"). By improving children's awareness of how words are built from smaller meaningful units (called *morphemes*), the intervention aimed to support more accurate and efficient word decoding.

Twelve high-frequency Italian derivational suffixes were selected based on their productivity (i.e., their ability to form many different words) and corpus frequency: *-ista, -tore, -mento, -zione, -aio, -eria, -aggio, -ismo, -ità, -oso, -bile, and -ale* [12]. A *derived word* is a word formed by combining a base with a suffix (i.e., *chitarra* "guitar" → *chitarrista* "guitarist"), whereas a *simple word* is a word that does not contain a derivational suffix (i.e., *chitarra*). Throughout the training, suffixes were introduced progressively following a gradient of *semantic complexity* (i.e., how abstract or conceptually complex the resulting word meaning is). This progression helped limit cognitive load and allowed children to focus on specific *morphemic units* (the meaningful components that form words). Agentive nominal suffixes (suffixes that create nouns referring to a person performing an action, i.e., *-ista* or *-tore*) were presented first, followed by suffixes forming *action nouns* (i.e., *-zione, -mento*) and *place nouns* (i.e., *-eria*, *-aio*). In later stages, the training introduced suffixes forming *abstract nouns* (i.e., *-ità*, *-ismo*) and *adjectives* (i.e., *-oso*, *-bile*, *-ale*). The progression across exercise sessions also considered syllabic length (4 or 5 syllables), as longer words are typically more demanding to decode for children with dyslexia. *Allomorphic variation* (i.e., cases in which the base word changes slightly in form when combined with a suffix while maintaining the same underlying meaning) was introduced only at advanced levels, when more transparent base–suffix mappings had been consolidated. For example, *fiore* → *floreale*, where the base form changes from *fior-* to *flor-* before the suffix *-ale*. For each exercise, a level comprised approximately 20 stimuli.

We developed 4 types of exercises to target complementary aspects of morphological processing, ranging from structural recognition to productive manipulation, relational mapping, and speeded decoding. In *Word Sort*, children were asked to classify words as simple or derived (i.e., *chitarra vs. chitarrista*) and place them in the corresponding basket. This exercise aimed to promote segmentation and sensitivity to derivational markers during decoding. *Fast Read* focused instead on speeded decoding: derived words were displayed for a controlled time interval (200-2200 ms) and had to be rewritten

after they disappeared. Exposure time was set by the therapist and adjusted according to the child's performance. It increased by 200 ms when overall accuracy fell below 80% and decreased otherwise, thereby promoting automaticity in the decoding without compromising its correctness. In *Word Building*, children had to combine bases and suffixes presented as cards to form existing derived words (i.e., *dente + ista = dentista*). This exercise encouraged children to reflect on the internal word structure and on the compatibility constraints between morphemic units (i.e., *dente + tore = dentatore* is an illegal word formation). *Word Puzzle* extended this analysis to morphological families. Children matched base and derived words sharing the same root, the core morpheme of a word that carries its primary meaning and cannot be further reduced, in order to clear the set of cards. This exercise encouraged family-level organization of the lexicon. In *Word Building* and *Word Puzzle*, a greater number of previously introduced suffixes could co-occur. This design choice intentionally increased task complexity, encouraged morphological reasoning when children attempted non-existing derivations, and provided cumulative reinforcement for the suffixes encountered in earlier sessions. The exercises were conceived as complementary components of a structured training path and were introduced progressively according to the child's linguistic level. While some activities were introduced later, they were designed to partially overlap with earlier ones (i.e., *Fast Read* and *Word Sort*) to reinforce the morphological processes at play.

2.1 Requirements

The design of the assistants was based on a set of core requirements defined during a workshop in collaboration with expert linguists specialising in developmental dyslexia, making sure that the LLM behaves as a specialized tutor rather than a generic chatbot.

The specific requirements for these assistants are as follows:

- Linguistic Knowledge: The assistant must have an advanced understanding of Italian morphology. It must accurately identify roots and affixes to provide scientifically sound explanations to the child.
- Context Awareness: The system must know the specific exercise being performed and the child's current state of progress to provide contextually relevant feedback and tracking.
- Child-Friendly Language: The assistant must communicate using simple, short sentences suitable for children. It must be able to translate complex linguistic concepts into accessible explanations.
- Proactivity: The assistant must be able to sense when a child is struggling and intervene autonomously with scaffolding or re-explanations to move the exercise forward.
- Affective Support: The interaction must account for the child's emotional state, providing encouragement during moments of frustration and celebrating successes to maintain high motivation.
- Companionship: The assistant must establish a sense of social presence, providing the child with a virtual avatar to interact with.
- Attention Management: The assistant must prevent topic drift. If the child attempts to discuss unrelated themes, the AI must gently but firmly redirect their attention back to the learning activity.

2.2 Prototype and Multimodal Interface

The prototype was implemented as a multimodal web interface designed to facilitate the interaction between the child and the assistant. To ensure the focus remains on morphological training rather than the burden of reading, the interface is split into two functional zones: an *Exercise Zone* on the left and a *Conversational Zone* on the right, where the actual representation of our companion Foxy is presented on top of the chat. Foxy is represented as an animated avatar capable of talking with the child and mimicking Ekman's six basic emotions [11] plus the idle, to increase engagement and empathy.

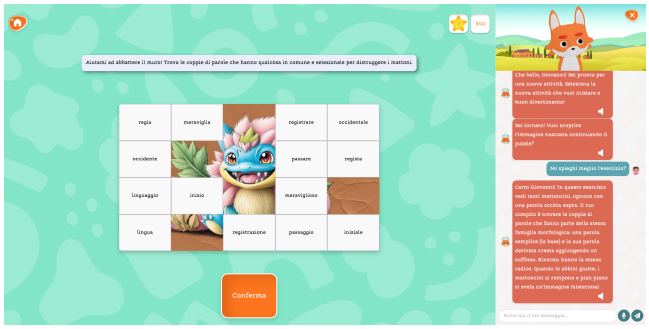


Figure 1: The multimodal interface. In the left area, the interactive linguistic exercise. On the right, the area for text-based or speech-based conversational interaction between the child and the virtual tutor Foxy

To accommodate the specific needs of children with dyslexia, the interface utilizes the *Leggimi* font [30], a typeface specifically designed for the Italian language to reduce character confusion and "letter crowding" effects. Interaction is inherently multimodal; the child can send messages via typing or talking to the microphone. The assistant's responses are delivered through a dual-output channel: they are displayed as text on the screen and simultaneously read aloud. This redundant output (audio-visual) allows the child to follow the text without being solely dependent on reading, thereby reducing cognitive load and listening effort.

2.3 Iterative Co-Design via Expert-Led Walkthroughs

The development of the assistants followed a participatory co-design involving four iterative cycles of testing and refinement. Central to this process was the use of Expert-Led Heuristic Walkthroughs [28].

Our approach utilized domain experts, specifically linguists, to systematically interact with the interface to identify obstacles in the UX-pedagogical flow. This method allowed for the identification of conceptual errors and "metalinguistic hallucinations" that a child might not be able to articulate, ensuring the system's behaviour aligned strictly with established therapeutic goals for dyslexia.

2.3.1 Procedure. Two expert linguists specialized in developmental dyslexia conducted the walkthroughs asynchronously over four development sprints. The procedure followed a three-step protocol:

(1) **Persona Simulation:** The experts engaged with the four exercise typologies (*Word Sort, Word Building, Fast Read,* and *Word Puzzle*) by simulating a variety of learners profiles, including "reluctant learners" (minimal or misleading input), "hyper-active learners" (excessive input), and "struggling learners" (repeated errors).
(2) **Documentation of Critical Points:** Experts compiled diagnostic reports for each iteration. These reports utilized annotated screenshots and conversation logs to pinpoint critical points, where the LLM's natural language generation conflicted with Italian morphological rules or the necessary emotional support.
(3) **Prompt Refinement:** Following each walkthrough, the qualitative feedback was synthesized to update the system instructions, specifically targeting the Foxy persona's tone, error-handling logic, and the precision of its metalinguistic explanations.

2.3.2 Mapping of Issues and Mitigations. Rather than isolated errors, the development process revealed recurring themes that evolved over the four iterations. Table 1 synthesizes these findings.

2.3.3 Discussion of the Process. The co-design process highlighted the necessity to balance the LLM's natural conversational capability and the type of dialogue required for speech therapy. Notably, some issues like "Metalinguistic Hallucination" re-emerged in later iterations as the complexity of the pedagogical instructions increased, requiring a more robust Lexical Knowledge Base within the prompt to constrain the Natural Language Generation (NLG).

In the early iterations (1 and 2), the focus was primarily on *conversational management*, preventing the model from being a general-purpose chatbot and ensuring it maintained its persona as an educational assistant. As the assistants became more stable, the focus in iterations 3 and 4 shifted toward *morphological accuracy*. A key breakthrough occurred in Iteration 4, where the concept of the "morphological family" was finalized. For example, by using the "heart" metaphor to describe the root of a word, the assistants were able to provide a visual and conceptual anchor for children with dyslexia, helping them see the relationship between base words and their derivatives (i.e., *lavoro* and *lavoratore*) more clearly.

By the final iteration, the assistants moved beyond simple correctness into a supportive guide that could offer descriptive feedback, encouraging the child to reflect on the structure of the words rather than focusing solely on the "right" or "wrong" answer.

3 Implementation and Prompt Engineering

3.1 Implementation

The prototype was implemented as a multimodal web interface inspired by the ChatCare [23] architecture.

The Speech-to-Text used to transcribe the user's voice was implemented using the *Web Speech API*. Text-to-Speech, instead, was delivered through *ElevenLabs API* for its performance in replicating human-like prosody and naturalness [5]. We utilized this state-of-the-art neural synthesis to provide a natural, high-fidelity Italian voice, which is essential as many standard TTS models lack the nuanced prosody required for the Italian language. Furthermore, the system integrates the *Emoty* emotion recognition model [9] to understand the emotion of the child and give more context to the LLM to make better and empathetic responses.

The core of the application's pro-activeness lies in its **Event-Based architecture**. Rather than only waiting for the child to ask for help, the system monitors the state of the exercise in real-time. Every action, such as selecting a morpheme or submitting an answer, is registered as an event that can trigger a state change in the conversational agent. For example:

- **Error-Driven Events:** If the system detects a pattern of repeated errors (i.e., five consecutive incorrect attempts), it triggers a "Proactive Nudge." This event injects a dynamic prompt, forcing the LLM to intervene with a scaffolding explanation.
- **Progress-Driven Events:** Upon reaching certain milestones (i.e., 50% completion), the application triggers a motivational event, prompting Foxy to provide affective support and celebrate small wins.

At the implementation level, *OpenAI API Assistants* are inherently reactive: they generate a response only after receiving a new message and cannot autonomously decide when to intervene. To satisfy the pro-activeness requirement, we therefore introduced a pragmatic workaround. When an event requires Foxy to intervene, the system adds a synthetic, labelled user message to the conversation thread (i.e., `THIS IS A PROACTIVE_ASSISTANT_QUESTION...`). This message does not correspond to an actual child request; instead, it acts as an internal trigger that provides the LLM with the current exercise state and prompts it to generate a proactive response. This solution should thus be interpreted as an implementation workaround, not as a deliberate architectural design choice.

3.2 Modular Prompt Structure

To ensure high-fidelity interaction and prevent cross-task interference, we implemented four distinct instances of the OpenAI Assistants API (specifically the `gpt-4.1` model). This architectural choice allows each instance to maintain a dedicated, pre-configured system prompt optimized for the specific mechanics of its exercise typology: *Word Sort, Word Building, Fast Read,* and *Word Puzzle*. By isolating these instructions into independent assistants, we eliminate the need to dynamically re-inject or modify core system instructions during a session, to ensure the model adheres strictly to the behavioural boundaries required for each specific morphological challenge.

The prompt sent to the OpenAI API is synthesized in real-time, combining **static instructions** (predefined pedagogical logic and persona) with **dynamic injections** (real-time application state). This synthesis ensures that while the core educational rules remain consistent, the assistants' responses are precisely grounded in the specific moment of the child's learning path.

To provide transparency regarding the final structure, the prompt template used for each API run follows this general pattern:

Table 1: Iterative Refinement of the AI assistants

Issue Category	Cycles	Problem Example	Mitigation Strategy
Topic Drift	1, 2	The assistants would engage in long off-topic discussions (i.e., about movies like *Lilo & Stitch*) if prompted by the child, even though they always tried to switch the focus back to the exercise.	Implemented "scaffolding stratagems" to acknowledge the child's comment but immediately redirect back to the task.
Terminology	1, 2	Use of overly simplistic language like "extra piece" or "added end" instead of formal linguistic terms.	Injected metalinguistic terms (*radice, suffisso*) paired with metaphors like the "heart of the word."
Metalinguistic Hallucination	1, 2, 4	The LLM occasionally invented non-existing words (i.e., *insegnatore* instead of *insegnante*).	Developed a verified Lexical Knowledge Base within the system instructions to restrict word generation.
Scaffolding Logic	3, 4	The LLM oscillated between being too helpful (giving answers) and too strict (refusing to give any examples).	Calibrated "hinting levels" to provide examples of *related* words before ever revealing the target word.
Error Feedback	2, 3	Simply stating an answer was "wrong" without explaining the morphological reason why.	Prompted the LLM to explain incorrectness in-depth (i.e., "your answer is wrong because these parts don't form a real word together").

```
The child you're talking to is called
<userFirstname> and they're <userGender>.
<baseInstructions> [...]
Based on the child's input, use this tone:
<emotionPrompt> [...]
Here below are the details about the exercise
that the child is doing: <exeStr>. [...]
The parts that the child already answered:
<exerciseActions>; The current progress of
the exercise: <exerciseProgress>;
<currentWordSection> [...]
<EventInstructions> [...]
<ExerciseExamples> [...]
```

The following sections detail how the highlighted parameters are structured and how they enable both context-awareness and pro-activeness.

3.2.1 **<userFirstname> <userGender>** . The prompt starts by defining the child's name and gender. While the child's name is important in building a connection with the user, the gender parameter is required because most Italian words change based on the person's gender. Unlike English, where a phrase like "you are good" remains the same for everyone, Italian adjectives and past participles must change their final letters (i.e., bravo for a boy vs. brava for a girl) to be correct. By providing this information upfront, the system ensures that Foxy avoids basic grammatical errors and maintains the linguistic precision necessary for an educational tutor.

3.2.2 **<baseInstructions>**. This block serves as the static foundational logic of the system, acting as the assistants' behavioural guide. It defines the persona, communication style, specific exercise knowledge, morphological knowledge, interaction rules, and conversational guardrails of the assistant.

Assistant Persona. The assistants are personified as Foxy, a peer-like, friendly tutor. The prompt explicitly defines the goal of the interaction: "`Your name is Foxy. You are an assistant whose objective is to help children aged 8 to 11 [...] to recognize the morphemes that make up words, in order to improve decoding in reading.`" By framing the identity around a specific developmental target, the system ensures that the LLM maintains its role as a specialized tutor rather than a generic conversationalist.

Communication Style. To accommodate the reduced cognitive processing speed and phonological difficulties associated with dyslexia, the system enforces strict communication rules: "`Use short and simple sentences, but make sure that all linguistic explanations are scientifically accurate and impeccable`". This part ensures that the child is not overwhelmed by "walls of text" while still receiving high-quality instruction. Furthermore, it grounds the AI in the physical reality of the application by instructing it to "`encourage the child to complete the exercise on the left side`" of the multimodal interface.

Specific Exercise Knowledge. This part of the prompt specifies to the assistant how the exercise works, explaining what task the child must complete and how the child interacts with that section of the screen. For example: `IMPORTANT: The user is playing the Fast Read game. This game is called Fast Read. This game helps improve attention and memory skills and works like this: A derived word will appear for a short time and then disappear, the child must remember and type the word from memory in the box, then a new word will appear, and the exercise continues`

Morphological Knowledge. A significant portion of the `<baseInstruction>` block is dedicated to the Lexical Knowledge Base that gives the LLM more context and information about morphological rules and avoids metalinguistic hallucinations (e.g. using

a non-existing word as an example). The system provides fixed definitions for every productive Italian suffix (i.e., "`-zione: nominal derivational suffix...`"). Furthermore, to handle the complexity of allomorphs without confusing the child, the prompt instructs the model to simplify technical jargon: "`When explaining to children, it is better not to use the term 'allomorph' and instead say that it is just a slightly different way to write the suffix.`"

Interaction Rules. The interaction logic is governed by strict negative constraints (no-spoiler rules): "`NEVER tell the child if their answer is correct or wrong if the question is still unanswered`". This forces the assistants to guide the child step-by-step rather than simply providing the solution.

Conversational Guardrails. Finally, to keep the session productive, a redirect mechanism is included: "`If child user asks unrelated questions [...] or wants to stop, kindly invite them to stay focused on the exercise`". This prevents topic drift and keeps the child engaged with the linguistic training path.

3.2.3 **`<emotionPrompt>`**. Once the system recognises the emotion of the user, it instructs the LLM on the specific tone to adopt based on the child's input. Below is presented the instruction after "sadness" is detected.

> "`Be empathetic and supportive. Acknowledge the user's feelings and offer comforting responses while maintaining an uplifting approach.`"

Additionally, the LLM is tasked with classifying the interaction into one of seven emotions (*joy, sadness, anger, fear, disgust, surprise, neutral*) for internal tracking of the conversation.

3.2.4 **`<exeStr> <exerciseActions> <exerciseProgress> <currentWordSection>`**. To provide context-aware assistance, the system injects the full exercise structure (`<exeStr>`), the history of user answers (`<exerciseActions>`), and the progress percentage (`<exerciseProgress>`). Furthermore, in some exercises, like *Fast Read* or *Word Sort*, the `<currentWordSection>` informs the assistants of exactly what item is currently visible on the screen, preventing the LLM from referencing future or previous parts of the exercise.

3.2.5 **`<EventInstructions>`**. When a series of events meets specific conditions, a dynamic prompt is injected via the `<EventInstructions>` block. A primary trigger for this proactive intervention is when a child makes five errors in a row. This threshold was specifically determined in consultation with domain experts based on the average length and item count of the exercises. The goal of this calibration is to balance autonomy and support: intervening too early could be perceived as frustrating or intrusive, while waiting too long might lead to total cognitive collapse. Upon activation, the parameter inserts the following instruction to trigger a conversational repair cycle:

> "`THIS IS A SPECIAL MESSAGE: The user answered wrong 5 times in a row, explain them what they did wrong by looking at the previous answers given.`"

This forces the assistant to draw the attention of the user, pause the exercise flow, and provide a personalized re-explanation based on the specific linguistic mistakes identified in the interaction history.

3.2.6 **`<ExerciseExamples>`**. To minimize LLM hallucinations, the system provides a specific set of verified examples, inside the `<example>` placeholder, via the following instruction:

> "`IMPORTANT: Here are the examples for you to use if the child asks for an example or if you want to explain the exercise by giving an example: <example>. Do not create your own example and don't take it directly from the exercise. USE ONLY THESE ONES. USE THEM ONLY ONE AT A TIME. If there are no examples available, tell the child you cannot give them examples for this level and re-explain the exercise.`"

By providing these specific references and a fallback logic for null-data states, the system ensures the assistants only validate or explain answers based on verified exercise examples, drastically reducing linguistic hallucinations.

4 Expert Evaluation

4.1 Goal and Research Variables

We conducted two exploratory pilot studies involving domain experts, namely educators and therapists, to gain insight from individuals who work daily with children. We evaluated the Foxy CA's perceived usability and usefulness, along with questions to understand how caregivers trusted the system in educational or therapeutic settings. The aim of these studies was to obtain a formative professional evaluation of the tailored assistants and collect refinement suggestions before conducting studies with children. Importantly, these studies were not intended to validate the system's clinical or educational efficacy, but rather to assess its preliminary acceptability and design soundness. The objectives of these studies are summarized in the following research question:
"To what degree is Foxy perceived as an engaging, usable, and useful "virtual tutor" for children during the execution of the dyslexia exercises?"

To answer the research question, a set of research variables and sub-variables (Table 2) has been identified and utilised to define the questionnaire later given to caregivers.
All the questionnaire items follow a 7-point Likert scale, where 1 indicates complete disagreement with the statement, 4 is the neutral point, and 7 indicates full agreement.

Table 2: Research variables and sub-variables with associated questionnaire questions

Research Variable	Research Sub-Variables	Survey Items
EN: **Engagement of Foxy**	EN1: **Attention**	EN1.1: Foxy captures the child's attention
		EN1.2: Foxy reduces the child's concentration on the exercise
		EN1.3: Foxy is NOT a distracting element for the child
	EN2: **Interaction Quality**	EN2.1: The interaction with Foxy is engaging for the child
		EN2.2: Foxy offers pleasant feedbacks for the children
		EN2.3: Interacting with Foxy is fun
US: **Usability of Foxy**	-	US1: Foxy is easy for children to use
		US2: Foxy sentences are understandable to the child
UT: **Utility of Foxy**	UT1: **Utility as Motivator**	UT1.1: Foxy increases the child's motivation to complete an exercise
		UT1.2: Foxy helps make the execution of the exercise less boring
		UT1.3: Foxy strengthens the child's motivation to do multiple exercises
	UT2: **Utility as Facilitator**	UT2.1: Foxy facilitates the execution of the exercises
		UT2.2*: Foxy facilitates the understanding of the exercises
	UT3: **Utility in Language Development**	UT3.1: Foxy provides appropriate feedback in case of repeated errors
		UT3.2: Foxy could promote the child's language development
LC: **Linguistic Correctness**	-	LC1: Foxy generates an acceptable number of linguistic errors (i.e misspellings)
		LC2: Foxy generates an acceptable number of conversational errors (i.e hallucinating or not properly answering)
FA: **Foxy Appropriateness**	-	FA1: Foxy's language style is appropriate
		FA2: Foxy's conversational flow is appropriate
		FA3: Foxy's feedback is appropriate and correct
		FA4: The emotions expressed by Foxy are appropriate to the conversational context
RK: **Risk**	-	RK1: Foxy should be used ONLY under adult supervision
		RK2: Foxy can be used without adult guidance
		RK3: Foxy could weaken the interaction with the caregiver

Table 3: Participant Demographics: Summary of Educators and Therapists

Characteristic	Educators ($n = 9$)	Therapists ($n = 9$)
Gender		
Male	0 (0%)	3 (33.33%)
Female	9 (100%)	6 (66.67%)
Age Range		
25-34	7 (77.78%)	3 (33.33%)
35-44	1 (11.11%)	3 (33.33%)
45-54	1 (11.11%)	3 (33.33%)

4.2 Participants

Participants were recruited on a voluntary basis. Therapists and educators were involved via schools and centres collaborating with us on various projects. To participate in the study, caregivers were required to have a degree and professional experience in educational or psychological fields strictly related to children. A total of nine educators and nine therapists were recruited for the study (Table 3).

4.3 Procedure

To each participant, four exercise sessions were assigned; the first session was **Word Sort**, followed by **Word Building**, **Fast Read**, and **Word Puzzle**. Each session consisted of four exercises of the same typology but with varying levels of difficulty (from easy to hard) for a total of 16 different exercises. The set of exercises was identical for all experts. Participants were asked to interact with Foxy to solve the exercises, trying to behave like their students/patients with dyslexia. Before starting the activities, caregivers received a brief presentation introducing LLMs, the application, and its intended purpose. After that, participants used the application independently, without any additional demonstration or guidance. After completing the activities, caregivers were asked to complete the customized questionnaire (Table 2).

5 Results

The results were analysed based on Mean (M) and Standard Deviation (SD) scores of the questionnaires' responses.

For items with a high variance (greater than or equal to 1.45), we have further inspected the values by considering the mean scores of therapists and educators and the corresponding SDs to understand if there were very different evaluations due to the diverse professional fields.

While we acknowledge that the small sample size reduce statistical strength and make the analysis explorative, we conducted Welch's t-test, to observe potential patterns and average differences between the two groups, while accounting for the small sample

size and potential inequality of variances. Statistical significance was set at $p < 0.05$.

5.1 Engagement of Foxy

The results showed that the experts had a neutral opinion regarding Foxy's impact on children's concentration, in not being a distracting element, and in its interactions being fun (EN1.2: M = 3.94, SD = 2.07; EN1.3: M = 4.27, SD = 2.02; EN2.3: M = 4.28, SD = 1.41).

For items EN1.2 and EN1.3, the variance scores were notably high, surpassing the 2.0 value. We further investigated the two items by analysing the mean scores and SDs by keeping the experts divided into educators and therapists. While teachers' mean scores (EN1.2: M = 4.44, SD = 2.07; EN1.3: M = 4.78, SD = 2.11) were higher than the therapists' ones (EN1.2: M = 3.44, SD = 2.07; EN1.3: M = 3.78, SD = 1.92), the variances kept being high in both groups.

In general, the experts slightly agreed that Foxy could capture children's attention (EN1.1: M = 5.33, SD = 1.24).

Divergent views were observed between educators and experts regarding Foxy's capturing children's attention (EN2.1: M = 4.94, SD = 1.59) and the pleasantness of its feedback (EN2.2: M = 4.67, SD = 1.5).

For EN2.1, experts' opinions were significantly different ($t(15.1) = 2.56$,$p = 0.022$); teachers considered the agent capable of engaging children (M = 5.78, SD = 1.20), while therapists had a neutral evaluation of it (M = 4.11, SD = 1.54), but with a high variability among their peers. Concerning the feedbacks, opinions were significantly ($t(13.8) = 3.79$, $p = 0.002$) more positive inside the teacher group (EN2.2: M = 5.67, SD = 0.87) than in the therapists case (EN2.2: M = 3.67, SD = 1.32).

5.2 Usability of Foxy

The results showed that the experts had on average a positive opinion regarding Foxy easiness of use (US1: M = 4.72, SD = 1.67) and understandability (US2: M = 4.83, SD = 1.29), but that at the same time in both cases significant differences were found (US1: $t(13.8) = 3.79$, $p = 0.002$; US2: $t(14.3) = 2.80$, $p = 0.014$). The educators (US1: M = 5.00, SD = 1.58) were more positive about Foxy being easy to use than the therapists (US1: M = 4.44, SD = 1.81), despite both groups presenting high variability. Educators were also significantly more favourable (M = 5.56, SD = 0.88) in evaluating Foxy as understandable to children than the therapists (M = 4.11, SD = 1.27).

5.3 Utility of Foxy

The results showed that the experts had a neutral opinion regarding Foxy's involvement in facilitating exercise execution (UT2.1: M = 4.39, SD = 1.29).

In general, caregivers slightly agreed that Foxy could facilitate the understanding of exercises (UT2.2: M = 4.95, SD = 1.39).

A statistically significant difference between therapists and educators emerged for UT2.2 ($t(13.2) = 3.13$, $p = 0.008$), with educators reporting significantly higher agreement than therapists regarding Foxy's potential in facilitating understanding of tasks.

Divergent views were observed between educators and experts regarding Foxy's effectiveness in reducing boredom (UT1.2: M = 5.11, SD = 1.45).

A closer examination of UT1.2 revealed that teachers' feedback (M = 5.67, SD = 1.22) was more positive compared to therapists' (M = 4.56, SD = 1.51), with the therapists' responses exhibiting greater dispersion within their group.

Experts' opinions diverged regarding Foxy's potential to support language development (UT3.2: M = 4.39, SD = 1.50), increase children's willingness to complete exercises (UT1.1: M = 5.11, SD = 1.45) or execute others (UT1.3: M = 5.28, SD = 1.49), and provide appropriate feedback in cases of repeated errors (UT3.1: M = 4.11, SD = 2.11).

The potential usefulness of the system in promoting language development, on average, achieved a neutral result; by analysing the two groups, teachers showed a positive opinion on it (UT3.2: M = 5.44, SD = 1.01), whereas therapists, on average, gave a slightly negative evaluation on the matter (UT3.2: M = 3.33, SD = 1.12). The difference in UT3.2 ratings between the two groups was statistically significant ($t(15.8) = 4.20$, $p<0.001$).

On average, teachers expressed strong agreement that the conversational agent could increase children's willingness to complete an exercise (UT1.1: M = 5.89, SD = 1.05) and engage in others (UT1.3: M = 6.11, SD = 0.78). In contrast, therapists adopted a more neutral stance (UT1.1: M = 4.33, SD = 1.41; UT1.3: M = 4.44, SD = 1.59), with considerable variability in their evaluations regarding the system's influence on users' willingness to engage in exercises. Differences between the two groups of experts on UT1.1 ($t(14.8) = 2.65$, $p = 0.019$) and UT1.3 ($t(11.7) = 2.82$, $p = 0.016$) where found to be significant. A statistically significant difference was also found for UT3.1 ($t(13.4) = 3.42$, $p = 0.004$). Teachers positively evaluated the appropriateness of the feedback in cases of repeated errors (M = 5.44, SD = 1.24), whereas therapists reported a negative evaluation (M = 2.78, SD = 1.98), with greater variability observed within the therapists' group.

5.4 Linguistic Correctness

Experts, on average, agreed on positively evaluating Foxy on the amount of linguistic and behavioural errors in its answers (LC1: M = 5.17, SD = 1.42; LC2: M = 5.11, SD = 1.32). Between the two groups, no significant differences were found.

5.5 Foxy Appropriateness

Caregivers, on average, agreed on the evaluation of Foxy's appropriateness. Experts perceived, on average, the language style as appropriate for the target profile (FA1: M = 4.72, SD = 1.23) and expressed emotions as consistent with the context (FA4: M = 4.72, SD = 1.41). Opinions about the flow of the conversation being appropriate for the users were generally neutral (FA2: M = 4.27, SD = 1.22), while the appropriateness and correctness of the feedback were slightly positive (FA3: M = 4.78, SD = 1.31).

5.6 Risk

Regarding RK scores, on average, RK1 (M = 3.28, SD = 1.36) and RK3 (M = 3.33, SD = 1.78) didn't reach the neutral value. For RK1, experts generally disagreed that adult supervision should be mandatory for children using Foxy. In contrast, RK3 results showed a disalignment concerning the possibility that the agent could weaken interactions between children and caregivers. Concerning the weakening of

Table 4: Mean and Standard Deviation scores of the questionnaire items. High SD are reported in bold, whereas items that were proven to be significant are indicated with *.

Item	Mean	SD	Item	Mean	SD	Item	Mean	SD
EN1.1	5.33	1.24	UT1.1*	5.11	**1.45**	LC2	5.11	1.32
EN1.2	3.94	**2.07**	UT1.2	5.11	**1.45**	FA1	4.72	1.23
EN1.3	4.27	**2.02**	UT1.3*	5.28	**1.49**	FA2	4.27	1.22
EN2.1*	4.94	**1.59**	UT2.1	4.39	1.29	FA3	4.78	1.31
EN2.2*	4.67	**1.50**	UT2.2	4.95	1.39	FA4	4.72	1.41
EN2.3	4.28	1.41	UT3.1*	4.11	**2.11**	RK1	3.28	1.36
US1*	4.72	**1.67**	UT3.2*	4.39	**1.50**	RK2	4.39	**1.46**
US2*	4.83	1.29	LC1	5.17	1.42	RK3	3.33	**1.78**

the interactions, teachers had a neutral opinion about it (RK3: M = 3.89, SD = 1,76), while therapists on average, didn't believe that the system could lessen the quality of the interactions (RK3: M = 2.78, SD = 1,72); both groups have a high variability on this matter among their own peers.

Experts' opinions varied concerning whether Foxy could be used without the guidance of an adult (RK2: M = 4.39, SD = 1.46), with educators (RK2: M = 4.89, SD = 1.36) being generally more open to it than therapists (RK2: M = 3.89, SD = 1.45), though no significant difference was found.

6 Discussion

6.1 Engagement of Foxy

In general, the experts slightly agreed that Foxy could capture children's attention, although their average rating regarding the extent to which Foxy provides fun interactions was neutral. Similarly, their assessment of Foxy's impact on attention was neutral, indicating uncertainty about whether Foxy might act as a distraction or negatively affect children's concentration. Concerning attention, there was high variability in responses, not only between therapists and educators but also within each group, reflecting substantial differences of opinion even among professionals from the same field. The main differences between educators and therapists concerned Foxy's capability of engaging children and the likeability of its feedback. In both cases, therapists' responses tended to be closer to neutral, slightly lower in the case of feedback, while educators were slightly positive. This divergence may reflect the different professional orientations of the two groups. Educators appeared to interpret Foxy from an engagement perspective more than therapists, who, in contrast, appeared to adopt a more clinically oriented perspective, giving greater weight to the potential clinical efficacy of Foxy and to its suitability as part of an intervention pathway than to its ability to attract the intended users.

6.2 Usability of Foxy

Overall, caregivers expressed a positive stance regarding Foxy's ease of use and understandability. On average, educators rated Foxy's usability significantly more positively than therapists did, particularly with respect to understandability. Therapists' evaluations, in contrast, were generally neutral.

Despite these overall trends, there was substantial variability in ratings within both groups. This dispersion indicates that not all caregivers perceived Foxy as easy for children to use, with some members in the groups reporting it to be difficult.

Educators' more positive assessment may also be explained by the similarity between Foxy's feedback style and the type of feedback teachers commonly provide during classroom exercises. In classroom contexts, feedback is often directed at supporting task completion, maintaining engagement, and guiding children through short explanations, hints, and encouragement. From this perspective, Foxy's interaction style may have been perceived as familiar and compatible with everyday educational practices. Moreover, the activities supported by Foxy were close to the types of morphology-oriented exercises that educators may already use in classroom training; therefore, teachers may have focused primarily on whether Foxy could provide useful procedural and explanatory support for carrying out the task. Therapists, instead, may have evaluated the system against a more individualized clinical standard. In therapeutic settings, the relationship between the clinician and the child is typically one-to-one, and feedback is expected to be highly tailored to the child's specific cognitive, linguistic, emotional, and relational profile. This may explain why therapists were more cautious in judging the appropriateness and usefulness of Foxy's feedback, particularly when considering its potential role within an intervention pathway.

6.3 Utility of Foxy

Foxy utility as a motivator was generally evaluated positively; its effectiveness in reducing boredom and strengthening children's motivation was, on average, deemed positive. However, educators and therapists differed significantly in their evaluations of Foxy's capacity to strengthen children's motivation to complete one or more exercises. Educators were very positive in believing the system was able to motivate children, with a low dispersion among peers, whereas therapists were neutral but with a higher dispersion in their evaluation.

Results showed that the experts had a neutral opinion regarding Foxy's role as Facilitator. Its involvement in facilitating exercise execution was evaluated in a neutral way, without showing differences between groups or high dispersion within them. Caregivers, on average, slightly agreed that Foxy could facilitate the understanding of exercises and exhibited a statistically significant difference

between therapists and educators. Educators reported significantly higher agreement than therapists regarding Foxy's potential in facilitating children's understanding of tasks.

Experts' opinions differed significantly regarding Foxy's utility for language development. Teachers found Foxy's feedback in cases of repeated errors to be appropriate and positively evaluated Foxy's potential usefulness in enhancing language skills. Therapists' opinions of Foxy's utility, by contrast, were slightly negative. In some cases, their evaluation of the appropriateness of the feedback was also negative, although responses showed high variability. The analysis of conversations from the two groups highlights differences in the topics addressed and in communication style, and sheds light on the different results emerging from the questionnaires. Educators tended to ask questions focused on linguistic constructs (i.e, "Why is *giornalista* [journalist] not a simple word?") to explore whether Foxy could support children's reflection on specific morphological structures. Morphological exercises are typically embedded in school-based literacy practices, and the educators in our study focused on Foxy's ability to support these practices. In contrast, therapists' interactions with Foxy focused more on evaluating the agent's relational capability instead of its morphological or educational competence. They often attempted to push the conversation beyond the scope of linguistic training and used prompts to "provoke" the conversational agent (e.g., "I don't like you!") or to move the dialogue onto a more emotionally or psychologically oriented basis (e.g., "Can you explain why you feel uncomfortable?"). Therapists' interest in Foxy's relational behaviour may reflect two factors: first, a different style of interaction with children during language training in clinical contexts compared with educational contexts; and second, therapists' lower familiarity with morphological training compared to educators. Indeed, this type of training, which is very frequent in school environments, is often not addressed in routine clinical practice, which tends to focus more extensively on phonological aspects.

6.4 Linguistic Correctness

Foxy's appropriateness was generally evaluated positively, with no significant differences between educators and therapists and no substantial variability within groups. This consistency suggests that professionals shared a broadly aligned opinion about Foxy's generating an acceptable number of linguistic and conversational errors.

6.5 Foxy Appropriateness

Caregivers' evaluations of Foxy's appropriateness were generally aligned, indicating a shared perception across professional groups. Overall, experts agreed that Foxy's communication style was suitable for the intended target population. In particular, the language used was rated as appropriate for the user profile, and the emotions expressed by Foxy were considered coherent with the interaction context. In contrast, evaluations of the conversational flow were more reserved. The appropriateness of the dialogue progression for users received a neutral mean rating, indicating some uncertainty regarding the naturalness or pacing of interactions. Finally, the appropriateness and correctness of Foxy's feedback were rated slightly positively. Taken together, these results point to an overall favourable opinion of Foxy's appropriateness, with strengths in language style and emotional congruence, and more moderate evaluations of conversational flow.

6.6 Risk

The overall pattern indicates a generally low perception of risk associated with Foxy, albeit with notable variability across items and professional groups. Experts generally disagreed with the assertion that adult supervision should be mandatory when children use Foxy. This finding indicates that the system was averagely perceived as sufficiently safe and manageable to allow for a degree of autonomous use.

Concerning the possibility that Foxy might weaken interactions between children and caregivers, the results revealed a more complex pattern. Overall, the mean score did not reach neutrality, indicating that professionals tended not to believe that the agent would undermine relational dynamics. However, differences emerged between groups. Educators expressed a neutral stance on this issue, whereas therapists more clearly disagreed with the notion that Foxy could diminish the quality of interactions. Both groups exhibited substantial internal variability, reflecting heterogeneous perspectives even among peers within the same profession.

Finally, opinions were more divided regarding whether Foxy could be used without adult guidance. Educators appeared, on average, more open to autonomous use compared to therapists. Although this difference did not reach statistical significance, it suggests, along with the other results, a trend toward greater caution among therapists.

Overall, these findings highlight a nuanced perception of risk, characterized by general confidence in the system's safety alongside variability in attitudes toward autonomy and relational impact.

6.7 Limitations

Despite Foxy's promising results for dyslexia training, several limitations bound the scope of our current findings.

First, this study relies on proxy evaluation. The evaluation relied on expert judgment to obtain preliminary insights prior to deployment with a sensitive target population. Although the 18 expert participants provided high-level pedagogical and clinical insight, caregivers acting as students cannot fully replicate the authentic cognitive load, phonological challenges, or spontaneous emotional responses of a child with dyslexia. Consequently, the findings should be interpreted as formative and preparatory rather than as evidence of user-level efficacy.

Second, although inferential statistical analyses were conducted to compare educators and therapists (independent-samples Welch's t-tests), the pilot nature of the study and the small sample size limit the statistical power and generalizability of the findings. The analyses should therefore be interpreted as exploratory. While statistically significant differences and effect sizes were examined, the results remain preliminary and indicative rather than definitive. The results represent an initial feasibility assessment and perception-based evaluation, not a definitive measure of the system's training efficacy.

Third, there are technical and transparency constraints inherent in using commercial LLMs. Operating as a "black box", the underlying model's reasoning process is non-transparent, and despite rigorous prompt engineering, the risk of unpredictable behaviour or minor hallucinations remains. Furthermore, while the event-driven architecture allows for a proactive tutor, the current implementation relies on a sequence of API calls (LLM generation, sentiment analysis, and neural TTS synthesis) that introduces a cumulative latency of approximately 2 seconds that could occasionally disrupt the natural flow of dialogue.

7 Conclusions

In this paper, we presented the design and preliminary evaluation of LLM-driven assistants tailored to support children with dyslexia through Italian morphological exercises. By implementing a modular prompt architecture and the "Foxy" persona, we developed a system aimed to provide proactive, context-aware scaffolding while maintaining strict linguistic accuracy.

The findings from our pilot expert-based evaluation indicate a cautiously positive assessment of Foxy, while also revealing some differences between educators and therapists. Engagement was rated as moderate, with uncertainty regarding its capacity to sustain attention or avoid distraction, and substantial variability within and across experts' groups. Usability was generally perceived positively, particularly by educators, whereas therapists' judgments were more neutral and heterogeneous. Foxy's motivational utility emerged as a key strength, especially for educators, who consistently endorsed its capacity to enhance children's engagement with tasks. Still, therapists expressed more reserved and dispersed evaluations, particularly concerning language development and feedback appropriateness. Linguistic correctness and communicative appropriateness were evaluated positively and consistently across groups, though conversational flow received more neutral ratings. Risk perception was generally low, with no strong concerns about safety or relational harm; however, therapists displayed greater caution regarding autonomous use.

Overall, Foxy achieved initial professional acceptability within sensitive educational and therapeutic contexts. However, its potential for adoption in real-life contexts remains conditional. Moving from a prototype to a clinically or educationally robust intervention will require evaluation measures that go beyond perceived usability and utility and assess actual training effectiveness.

Furthermore, the linguistic content of the exercises and Foxy's behaviour are strongly tied to Italian, and specifically to Italian derivational morphology. The training supported by Foxy focuses on morphological decomposition by promoting the identification of derivational suffixes, base–suffix segmentation, morphological families, and allomorphic variation relevant to word-level decoding. These linguistic targets are necessarily language-specific and would need to be redesigned for other languages, taking into account their phonological and morphological properties. Consequently, the behaviour of the conversational assistant would also need to be adapted. What is transferable and reusable is the implementation architecture and the design methodology for the conversational tutor, which are based on a verified knowledge base, a closed set of examples, topic-drift control, scaffolded feedback, event-driven pedagogical logic for proactive interventions, and multimodal text/audio support. In other languages, the same approach could be applied to address language-specific morphological training.

From a broader perspective, our research suggests that the sustainable advancement of AI-supported training intervention in sensitive educational and therapeutic settings will depend not only on technical improvements but on interdisciplinary participatory design and systematic evaluation that safeguard the centrality of human expertise. However, the function of these digital assistants must be considered as augmentative rather than a replacement for trained practitioners.

8 Future Work

The most immediate next step involves longitudinal user testing. Direct experimentation with children diagnosed with developmental dyslexia, tracking of morphological awareness and decoding performance, and comparative studies against traditional interventions will be essential to determine whether proactive LLM-based scaffolding produces statistically significant benefits. To this end, study protocols have been defined and and a cohort of participants has already been recruited.

Future work will also address the extension of Foxy's scope and its technological enhancement. The current version of Foxy was integrated only in exercises addressing morphological competence. Still, the training platform in which Foxy is embedded, designed to enhance reading skills in Italian children with dyslexia, is being progressively extended to address both phonological and morphological abilities, two important dimensions of decoding skills. We will extend Foxy's capabilities so that it can also act as an interactive conversational tutor during phonological training. Furthermore, to address the limitations of commercial APIs, we plan to transition to a local LLM deployment on a dedicated server and explore streaming architectures. This move aims to significantly reduce latency, ensure higher data privacy for sensitive user groups, and eliminate the long-term economic costs associated with third-party API dependencies.

Finally, while the current version of Foxy is tailored to the specificities of Italian morphology, our modular architecture is extensible and reusable, and our future work will explore adapting our approach to other languages.

9 Generative AI Use Disclosure

Generative AI tools, including Gemini 3.1 Pro and ChatGPT 5.2, were utilized solely to improve the readability, grammar, and flow of the manuscript. After using these tools, the author(s) thoroughly reviewed and edited the content as needed, and take(s) full responsibility for the final content and integrity of the publication.

Acknowledgments

This work was partially supported by: i) the Italian National Recovery and Resilience Plan (PNRR), Mission 4, Component 2, Investment 1.1,funded by the European Union – NextGenerationEU, under the PRIN 2022 project "A Technology-driven linguistic intervention for children with lan-guage and literacy weaknesses", CUP:

B53D23014380006; ii) Project DIMMI!- Programma Regionale FESR 2021-2027, Asse 1 – Obbiettivo specifico 1.1 – Azione 1.1.3 - Bando "COLLABORA & INNOVA" - CUP: E49I25001410007.

References

[1] Stephanie Al Otaiba, Amy Gillespie Rouse, and Kristi Baker. 2018. Elementary grade intervention approaches to treat specific learning disabilities, including dyslexia. *Language, Speech, and Hearing Services in Schools* 49, 4 (2018), 829–842.

[2] V. Alevizos et al. 2024. Handwriting Anomalies and Learning Disabilities through Recurrent Neural Networks and Geometric Pattern Analysis. *arXiv preprint* (2024). https://arxiv.org/abs/2405.07238

[3] Ann W Alexander and Anne-Marie Slinger-Constant. 2004. Current status of treatments for dyslexia: Critical review. *Journal of child neurology* 19, 10 (2004), 744–758.

[4] Merve Altındağ and Ayşe Kumaş. 2025. Evaluating a Phonological Awareness Intervention (PAI) Game for Children with Dyslexia. *British Journal of Educational Technology* (2025). doi:10.1111/bjet.13245

[5] Linda Bakkouche, Charles McGhee, Emily Lau, Stephanie Cooper, Xinbing Luo, Madeleine Rees, Kai Alter, Brechtje Post, and Julia Schwarz. 2025. Finding the Human Voice in AI: Insights on the Perception of AI-Voice Clones from Naturalness and Similarity Ratings. In *Interspeech 2025*. Rotterdam, The Netherlands, 2190–2194. doi:10.21437/Interspeech.2025-947

[6] Eleonora Aida Beccaluva, Fabio Catania, Fabrizio Arosio, and Franca Garzotto. 2024. Predicting developmental language disorders using artificial intelligence and a speech data analysis tool. *Human–Computer Interaction* 39, 1-2 (2024), 8–42.

[7] Sara Bertoni, Chiara Andreola, Sara Mascheretti, Sandro Franceschini, Milena Ruffino, Vittoria Trezzi, Massimo Molteni, Maria Enrica Sali, Antonio Salandi, Ombretta Gaggi, Claudio Palazzi, Simone Gori, and Andrea Facoetti. 2024. Action video games normalise the phonemic awareness in pre-readers at risk for developmental dyslexia. *npj Science of Learning* 9 (2024), 25. doi:10.1038/s41539-024-00230-0

[8] A. Castles and M. Coltheart. 2004. Is there a causal link from phonological awareness to success in learning to read? *Cognition* 91 (2004), 77–111.

[9] Fabio Catania, Nicola Di Nardo, Franca Garzotto, and Daniele Occhiuto. 2019. Emoty: An Emotionally Sensitive Conversational Agent for People with Neurodevelopmental Disorders. In *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*. University of Hawai'i at Manoa, 2014–2023. doi:10.24251/HICSS.2019.245

[10] Fabio Catania, Micol Spitale, Giulia Cosentino, and Franca Garzotto. 2021. Conversational Agents to Promote Children's Verbal Communication Skills. *Chatbot Research and Design. CONVERSATIONS 2020* (2021), 158–172. doi:10.1007/978-3-030-68288-0_11

[11] Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion* 6, 3-4 (1992), 169–200. doi:10.1080/02699939208411068

[12] Livio Gaeta and Davide Ricca. 2003. Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Rivista di Linguistica* 15, 1 (2003), 63–98.

[13] S. M. Goodman et al. 2024. AI Writing Assistance for Adults with Dyslexia Using Large Language Models. *Commun. ACM* (2024). doi:10.1145/3626952

[14] Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing* (3rd ed.). Pearson. Draft available online.

[15] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[16] J. L. Keelor, N. A. Creaghead, N. H. Silbert, A. D. Breit, and T. Horowitz-Kraus. 2023. Impact of text-to-speech features on the reading comprehension of children with reading and language difficulties. *Annals of Dyslexia* 73, 3 (2023), 469–486. doi:10.1007/s11881-023-00281-9

[17] K. C. Levesque and S. H. Deacon. 2022. Clarifying links to literacy: How does morphological awareness support children's word reading development? *Applied Psycholinguistics* 43, 4 (2022), 921–943. doi:10.1017/S0142716422000194

[18] Subhankar Maity and Aniket Deroy. 2024. Generative AI and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650* (2024).

[19] Q. T. Nguyen et al. 2025. Learning Disorder Detection Using Eye Tracking. *Proceedings of ACM Conference on Human Factors in Computing Systems* (2025). doi:10.1145/3715669.3726785

[20] Orton Society. 1995. Definition of dyslexia: report from committee of members. *Perspectives* 21 (1995), 16–17.

[21] A. Paglialunga et al. 2025. The Effectiveness of Artificial Intelligence-Based Interventions for University Students with Dyslexia. *Brain Sciences* 15, 8 (2025), 806. doi:10.3390/brainsci15080806

[22] Robin L Peterson and Bruce F Pennington. 2012. Developmental dyslexia. *The lancet* 379, 9830 (2012), 1997–2007.

[23] Francesco Piferi, Giovanni Caleffi, Giulia Valcamonica, Pietro Crovari, and Franca Garzotto. 2025. Integrating Large Language Models into Therapeutic Education for Children with Dyslexia: a Multimodal Framework. In *Proceedings of the D-SAIL Workshop-Transformative Curriculum Design: Digitalisation, Sustainability, and AI Literacy for 21st Century Learning, CEUR-WS, Palermo, Italy*. 70–79.

[24] Shen Qiao, Samuel Kai-Wah Chu, and Susanna Siu-sze Yeung. 2023. Understanding how gamification of English morphological analysis in a blended learning environment influences students' engagement and reading comprehension. *Computer Assisted Language Learning* (2023), 1–29. doi:10.1080/09588221.2023.2230273 Advance online publication.

[25] Sandra Raffoul and Lindsey Jaber. 2023. Text-to-speech software and reading comprehension: The impact for students with learning disabilities. *Canadian Journal of Learning and Technology* 49, 2 (2023), 1–18.

[26] Maria Rauschenberger, Ricardo Baeza-Yates, and Luz Rello. 2025. Screening Dyslexia Using Visual Auditory Computer Games and Machine Learning. *IEEE Access* 13 (2025), 29541–29553. doi:10.1109/ACCESS.2025.3539719

[27] Dan Reynolds and Amanda Goodwin. 2016. Supporting students reading complex texts: Evidence for motivational scaffolding. *Aera Open* 2, 4 (2016), 2332858416680353.

[28] Andrew Sears. 1997. Heuristic walkthroughs: Finding the problems without the noise. *International journal of human-computer interaction* 9, 3 (1997), 213–234.

[29] D. Shankweiler, S. Crain, L. Katz, et al. 1995. Cognitive profiles of reading-disabled children: comparison of language skills in phonology, morphology and syntax. *Psych Sci* 6 (1995), 145–156.

[30] Sinnos. 2006. *Leggimi!* https://www.sinnos.org/leggimi/le-font/

[31] M. J. Snowling. 1995. Phonological processing and developmental dyslexia. *J Res Read* 18 (1995), 132–138.

[32] Manuel Soriano-Ferrer and Manuel Morte-Soriano. 2017. Teacher perceptions of reading motivation in children with developmental dyslexia and average readers. *Procedia-Social and Behavioral Sciences* 237 (2017), 50–56.

[33] Marta Tagliani, Maria Vender, Giulia Valcamonica, Giovanni Caleffi, Franca Garzotto, and Denis Delfitto. 2025. DEA – An Innovative Technological Tool for Personalized Linguistic Training for Italian Children with Developmental Dyslexia. In *Proceedings of AIUCD 2025*. https://aiucd2025.dlls.univr.it/assets/pdf/papers/93.pdf

[34] Julien Vonthron, Claire Dupont, and Pierre Hubert. 2024. Rhythmic Training for Children with Dyslexia: Evidence from the Mila-Learn Serious Game. *Journal of Computer Assisted Learning* (2024). doi:10.1111/jcal.12789

[35] Evdokia Voultsiou and Lefteris Moussiades. 2025. A systematic review of AI, VR, and LLM applications in special education: Opportunities, challenges, and future directions. *Education and Information Technologies* 30 (2025), 19141–19181. doi:10.1007/s10639-025-13550-4

[36] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. arXiv:2403.18105. https://arxiv.org/abs/2403.18105

[37] Liang Xu, Monica Ward, Jenny Thomson, and Elaine Uí Dhonnchadha. 2024. Enhancing language learning for dyslexic learners: Integrating text-to-speech AI in CALL. *CALL for Humanity - EuroCALL 2024 Short Papers* (2024). https://api.semanticscholar.org/CorpusID:274905179

[38] H. Zoppé et al. 2025. Early detection of neurodevelopmental disorders in young children. *Early Human Development* (2025). doi:10.1016/j.earlhumdev.2025.106405