Andre Luis Lessa Junior

Multimodal LLMs for Augmentative Communication: Generating Story-Based Pictogram Boards from Images of Drawings, Fictional Characters, and Real-World Scenes

SÃO PAULO
2025

Andre Luis Lessa Junior

Multimodal LLMs for Augmentative Communication: Generating Story-Based Pictogram Boards from Images of Drawings, Fictional Characters, and Real-World Scenes

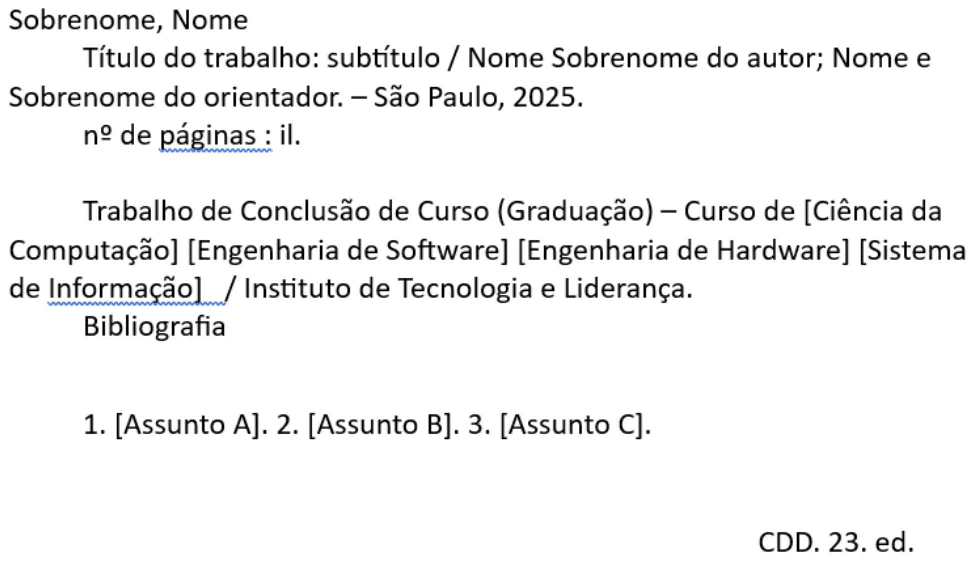
Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Computer Science

Advisor: Prof. Ms. Tomaz Mikio Sasaki

Coadvisor: Prof. Dra. Marilia Bense Othero

SÃO PAULO
2025

Cataloging in Publication
Library and Documentation Service
Instituto de Tecnologia e Liderança (INTELI)
Data entered by the author.



Sobrenome, Nome
Título do trabalho: subtítulo / Nome Sobrenome do autor; Nome e Sobrenome do orientador. – São Paulo, 2025.
nº de páginas : il.

Trabalho de Conclusão de Curso (Graduação) – Curso de [Ciência da Computação] [Engenharia de Software] [Engenharia de Hardware] [Sistema de Informação] / Instituto de Tecnologia e Liderança.
Bibliografia

1. [Assunto A]. 2. [Assunto B]. 3. [Assunto C].

CDD. 23. ed.

Acknowledgments

I would like to express sincere gratitude to Professor Marilia Bense Othero and Professor Tomaz Mikio Sasaki for their guidance and support throughout this research. Their insights were fundamental to refining the understanding of the research problem and shaping the methodological and conceptual directions of this work. I also thanks all the research participants for generously dedicating their time and expertise, making this study possible and contributing meaningfully to the advancement of AAC-oriented research.

Resumo

Lessa, Andre. **LLMs Multimodais em Comunicação Aumentativa**. 2025. nº de folhas. TCC (Graduação) – Curso Ciência da Computação, Instituto de Tecnologia e Liderança, São Paulo, 2025.]

Os sistemas de Comunicação Aumentativa e Alternativa (CAA) apoiam crianças com necessidades complexas de comunicação; no entanto, as pranchas tradicionais de pictogramas exigem uma ampla personalização manual. Embora estudos recentes tenham explorado a geração automatizada de vocabulário por meio de visão computacional e processamento de linguagem natural, o potencial dos Modelos de Linguagem de Grande Escala multimodais (LLMs) para a criação de materiais de CAA centrados na criança permanece amplamente inexplorado. Este estudo propõe uma abordagem multimodal baseada em LLMs para gerar automaticamente histórias educativas curtas e pranchas de pictogramas correspondentes a partir de imagens de personagens, desenhos e fotografias do mundo real, avaliando sua potencial contribuição para práticas comunicativas e terapêuticas voltadas a crianças autistas não verbais.

Palavras-Chave: Inteligência Artificial; Comunicação Aumentativa e Alternativa (CAA); Sistema de Comunicação por Troca de Figuras (PECS); Inteligência Artificial Multimodal; Tecnologia Assistiva

Abstract

Lessa, Andre. **Multimodal LLMs for Augmentative Communication**. 2025. nº of pages. Final course project (Bachelor) – Course Computer Science, Institute of Technology and Leadership, São Paulo, 2025.

Augmentative and Alternative Communication (AAC) systems support children with complex communication needs, however traditional pictogram boards require extensive manual customization. While recent studies have explored automated vocabulary generation using computer vision and natural language processing, the potential of multimodal Large Language Models (LLMs) for creating child-centered AAC materials remains largely unexplored. This study proposes an LLM-based multimodal approach to automatically generate short educational stories and corresponding pictogram boards from images of characters, drawings, and real-world photographs, and evaluates its potential contribution to communication and therapeutic practices for non-verbal autistic children.

Key words: Artificial Intelligence, Augmentative and Alternative Communication (AAC), Picture Exchange Communication System (PECS), Multimodal AI; Assistive Technology

List of Illustrations

List of Tables

List of Abbreviations and Acronyms

AAC 1 Augmentative and Alternative Communication 1

LLM 2 Large Language Model 2

PECS 3 Picture Exchange Communication System 3

Summary

1 Introduction

Augmentative and Alternative Communication (AAC) systems play a crucial role in supporting children with complex communication needs and are widely recognized for enabling participation in educational, social, and daily activities. However, fully harnessing the benefits of AAC technologies often requires substantial effort from families, professionals, and the users themselves. According to Waller et al, this burden is frequently intensified by poor usability, high learning demands, limited professional expertise, and difficulties in physical access.

Within AAC systems, one of the most widely adopted approaches is the Picture Exchange Communication System (PECS), which relies on pictograms to help non-verbal children express actions, needs, and preferences (Charlop-Christy et al.). Although effective, PECS demands continuous personalization, and many communicative situations fall outside the child's predictable routine. As a result, caregivers and professionals must invest significant time and effort to create and maintain appropriate communication boards.

As highlighted by Di Paola et al. AAC is not a one-size-fits-all solution and must be adapted to each child's interests, developmental profile, and communicative environment. This personalization challenge becomes even more complex when children prefer non photographic content such as drawings, sketches, or cartoon characters materials that traditional AAC tools and computer vision pipelines frequently struggle to interpret.

To address these limitations, this exploratory study investigates whether multimodal Large Language Models (LLMs) can overcome constraints observed in prior computer vision–based approaches to AAC content generation. Studies such as Vargas et al. highlight limitations in the semantic interpretation of drawings and fictional characters by traditional vision models. In contrast, this work explores how multimodal LLMs can generate short, child-centered stories and corresponding pictogram boards from images of drawings, fictional characters, and real-world photographs. By integrating visual understanding with language generation, the proposed approach examines the potential of next-generation AI models to support more flexible and engaging AAC materials for non-verbal children.

2 Development

2.1 Discover and Understanding Solution

The first module of a research project that investigates the potential of Artificial Intelligence to support the communication of nonverbal autistic children through Augmentative and Alternative Communication (AAC). The module focuses on understanding the problem space, existing solutions, and innovation opportunities, while outlining the conceptual and methodological foundations of the proposed system. Across five sprints, the project involved a literature review, market analysis, feasibility assessment, and social impact evaluation, leading to the definition of a scalable solution architecture. The study builds on prior work in automated AAC content generation, initially considering computer vision approaches but ultimately adopting a multimodal Large Language Model (LLM) due to its broader contextual understanding and flexibility. In addition, discussions with academic experts and institutions informed the evaluation strategy and potential expansion to other populations, such as children with cerebral palsy. Overall, Module 1 establishes the theoretical, technical, and social groundwork for subsequent development phases, emphasizing accessibility, inclusion, and the role of AI in enhancing communication and educational participation.

2.2 Application Development

The second module of the project, focused on the development and initial validation of the proposed AAC application. During this phase, the system architecture was implemented with a React Native front end and a FastAPI back end, integrating the ARASAAC API for pictogram retrieval and the Gemini model for story generation. The development process included experiments with a proprietary image recognition model based on BLIP; however, initial results revealed limitations related to dataset choice and training configuration, motivating the planned creation of a new, domain-specific image database. In parallel, the application underwent expert validation with an academic specialist, which confirmed high accuracy in pictogram identification while highlighting necessary adjustments related to content suitability and user experience. By the end of Sprint 2, the project delivered a functional minimum viable product (MVP), establishing a solid technical foundation for further refinement, ethical review, and future improvements to the image recognition pipeline and overall system usability.

2.3 Testing Application

The module presents the testing and validation phase of the AAC application, focusing on the evaluation of its technical infrastructure, AI models, and preliminary usability outcomes. During this stage, the system was deployed in a cloud-native architecture combining a React-based frontend, a FastAPI backend, and scalable services on Google Cloud Run, with secure data management through PostgreSQL and Google Cloud Storage. Two multimodal AI models were integrated and compared: the open-source LLaVA Next, selected for reproducibility and transparency, and the Gemini API, adopted for its superior performance and scalability. Initial testing with therapists and educators indicated satisfactory semantic alignment of generated pictograms, while highlighting the need for further prompt optimization to improve narrative coherence in the generated stories. Although testing with children is still pending ethical approval, professional feedback emphasized the application's potential for therapeutic and educational contexts. Overall, Module 3 establishes the system's technical robustness and feasibility, while identifying clear directions for refinement, ethical compliance, and future field validation.

2.4 Validation, Ethics Approval, and Submission

Module 4 represents the submission and validation phase of the project, consolidating both empirical evaluation and academic dissemination. During this stage, testing sessions were conducted with nonverbal autistic children to validate the system's results in real educational and therapeutic contexts, allowing the assessment of engagement, usability, and practical applicability of the generated pictograms and stories. In parallel, all pending documentation was submitted to the Ethics Review Board, resulting in formal approval to conduct and report the study.

Based on the outcomes of the testing phase, the first complete version of the scientific article was written and presented to an academic examination committee for feedback and refinement. Finally, the revised manuscript was submitted to *Communications of the ACM*, marking the transition from system development and testing to scholarly contribution and dissemination within the computing research community.

2.5 Methodology

The study adopted an exploratory research design aimed at investigating how artificial intelligence can assist therapists, teachers, and caregivers in the rapid creation of context specific communication boards. The proposed system automatically generates short, child appropriate stories and the corresponding set of pictograms directly from an input image, resulting in a communication board tailored to the visual context presented.

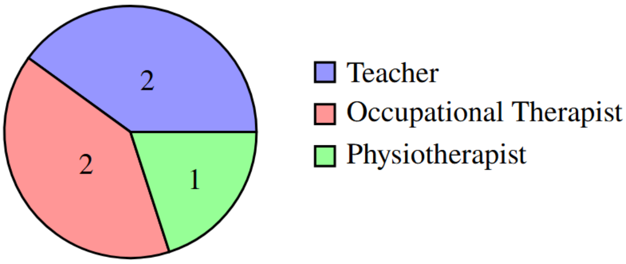
2.5.1 Participants

To evaluate the system, two complementary approaches were used, a quantitative assessment with professionals experienced in AAC and a qualitative observation of children interacting with the generated boards.

2.5.2 Professionals

The professional sample consisted of individuals who work directly with non-verbal autistic children or regularly employ communication boards as part of their clinical or educational practice. In total, five professionals participated in the study, distributed in the following roles:

Figure 1. Distribution of professional participants (n = 5).



To analyze the perceptions of professionals, the responses to the questionnaire were organized into four evaluative categories: pictogram and story quality, usability and system performance, therapeutic impact and integration and professional applicability. All elements were rated using a four-point Likert scale ranging from 1 (insufficient) to 4 (fully satisfactory). The complete set of questionnaire items is presented below.

- Q1 How do you rate the quality of the pictograms presented?
- Q2 How do you rate the clarity and adequacy of the generated stories?
- Q3 The application interface is easy to understand and use.
- Q4 The clarity of captions and pictograms meets the needs of clinical practice.
- Q5 The system’s response time is adequate for the therapeutic context.
- Q6 The tool helped increase the child’s interest and attention during the session.
- Q7 The resource facilitates the child’s communication or expression.
- Q8 The application reduced signs of frustration or interaction difficulties.
- Q9 The resource complements other AAC strategies already used in therapy.
- Q10 Overall, I consider the tool useful for therapeutic work.
- Q11 I believe this resource can generate long-term benefits for the child’s communication.

The correspondence between categories and questionnaire items is summarized in Table 1.

Table 1. Categories and corresponding questionnaire items.

Category	**Included Questions**
Pictogram and Story Quality	Q1, Q2

Usability and System Performance	Q3, Q5
Therapeutic Impact	Q4, Q6, Q7, Q8
Integration and Professional Applicability	Q9, Q10, Q11

2.5.3 Non-Verbal Children

The study included a limited number of sessions with children aged 5 to 8 years, which does not permit long-term conclusions regarding therapeutic outcomes. Nevertheless, these sessions provided valuable insights into the children's immediate engagement with the platform. The analysis focused on indicators such as duration of interaction, interest in exploring new images, responsiveness to auditory feedback from the pictograms, and spontaneous attempts to construct phrases using the generated vocabulary.

To ensure relevance and familiarity, each session began by asking the responsible adult which characters, drawings, or activities the child particularly enjoyed. An image corresponding to these preferences was then uploaded into the system. After the communication board was generated, the researcher initiated interaction by tapping on pictograms and letting the child listen to the synthesized speech. The child was encouraged but not instructed to explore the board, repeat sounds, and eventually combine pictograms into simple phrases if interested. This approach allowed the sessions to capture naturalistic engagement rather than task driven performance.

The child-focused evaluation was conducted in two formats. First, two autistic non-verbal children participated in supervised sessions at the Laboratório de Estudos sobre Infância e Deficiência (LEIA) at the University of São Paulo (USP), where researchers directly observed their interaction with the platform.

Additionally, one teacher independently tested the tool with their own non-verbal students during classroom activities, using photos of the child in real contexts such as holding a water bottle or painting a drawing rather than cartoon characters or external images.

These personalized inputs allowed the evaluation to capture how the system performed when processing daily situations that are meaningful to the child. This external testing offered complementary insights into the platform applicability and its potential integration into real educational and therapeutic routines.

2.5.4 Procedures

The procedures adopted in the study reflect the two distinct participant groups described above.

- Professionals accessed the platform, explored its functionalities, and completed the evaluation questionnaire.
- Children interacted with the tool during supervised sessions.
- A teacher independently tested the platform with a non-verbal student and later provided feedback.

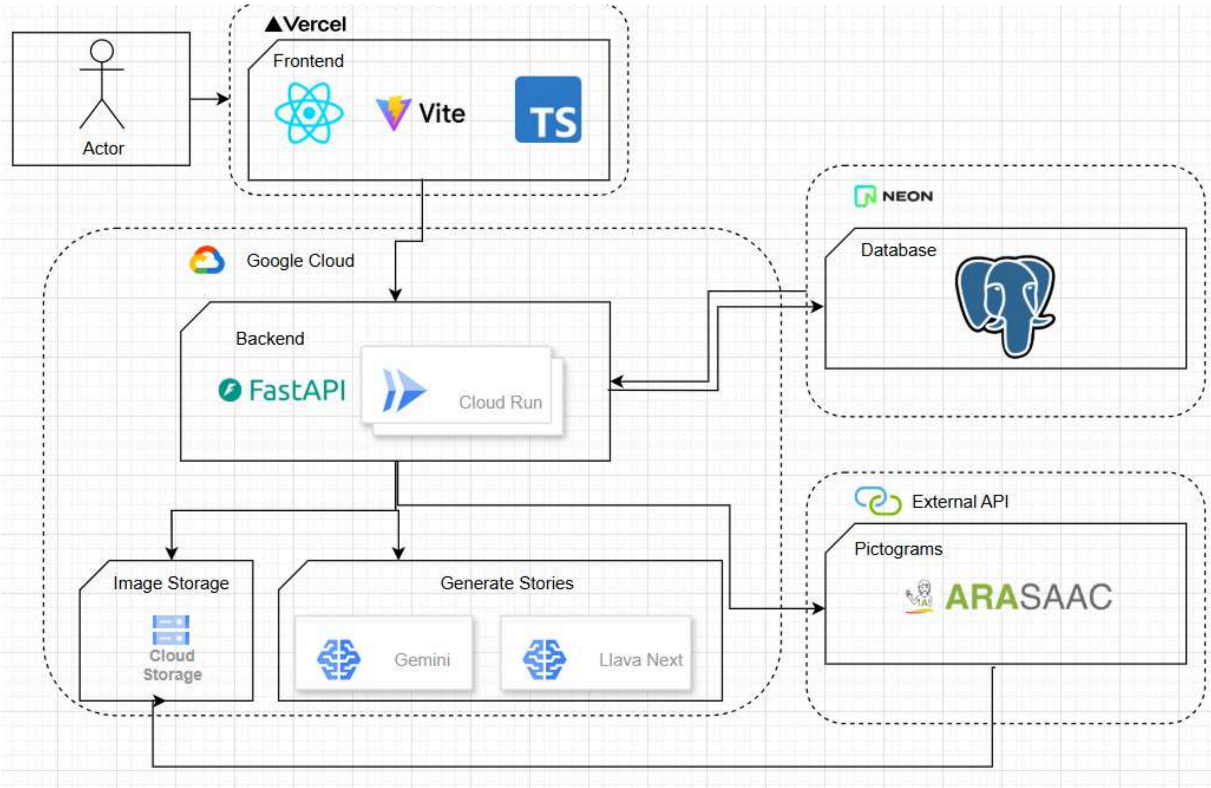
2.5.5 Data Analysis

Quantitative data were analyzed by computing the mean score within each category of the questionnaire. Qualitative data were examined through observational analysis of children’s behavior, engagement, and independence during interaction with the communication boards.

2.5.6 System Design

The system was designed as a modular architecture composed of three main components: the front-end, the back-end, and the artificial intelligence processing layer. This modularization enables independent development, clear separation of concerns, and easier maintenance testing.\newpage The infrastructure was built with cost-efficiency in mind, with the only operational expense being GPU runtime on Google Colaboratory Pro+ (USD 50) for running the LLaVA-Next model during experimentation.

Figure 2. System architecture diagram.



2.5.7 Front-End Module

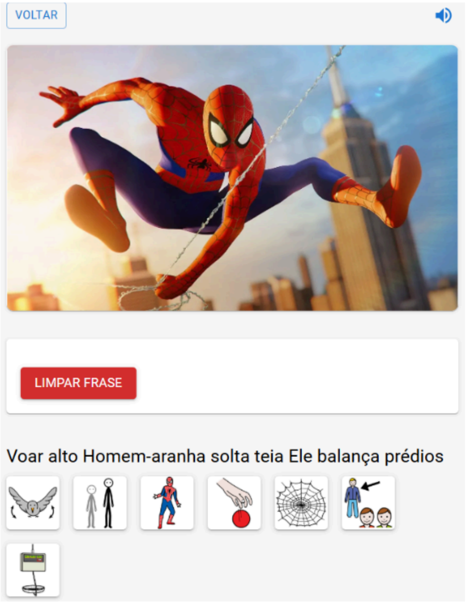
The front-end module was designed to provide a minimal, intuitive, and accessible user experience, adopting a fully responsive layout to ensure consistent usability across mobile devices, tablets, and desktop environments. Its implementation in TypeScript and React enables component modularity, predictable state management, and improved maintainability throughout the development lifecycle.

User authentication is supported through a dedicated login interface secured with a JSON Web Token (JWT) mechanism, ensuring that all generated communication boards and user albums remain protected. A registration interface is also included, containing a simplified form that collects only essential credentials (email and password), following good practices of minimal data handling.

From an architectural perspective, the front-end is organized into four primary pages:

- **Login/Registration Page** - Responsible for user authentication and account creation, relying on a secure JWT-based process.
- **Home Page** - Enables image acquisition from the device's gallery or camera and provides navigation to the album interface.
- **Album Page** - Displays all communication boards previously generated by the authenticated user.
- **Board Page** - Presents the generated board, including the original uploaded image, the pictograms retrieved from the ARASAAC API, and the corresponding short stories produced by the AI model (Figure 3).

Figure 3. Board Page interface used during user testing.



Voar alto Homem-aranha solta teia Ele balança prédios

For performance optimization, the system was built using the Vite bundler, which provides fast compilation, efficient hot-reload capabilities, and reduced build times. Deployment was carried out through Vercel, a cloud hosting platform that offers seamless integration with GitHub repositories and cost-free hosting for front-end applications, supporting continuous delivery and automated version control.

2.5.8 Back-End Module

The back-end module functions as the core orchestration layer of the system, mediating communication between the front-end interface, the artificial intelligence models, the ARASAAC pictogram API, and the cloud storage services. It was developed using FastAPI, a high-performance Python framework designed for RESTful architectures, enabling clear route organization, strong typing, and efficient request handling.

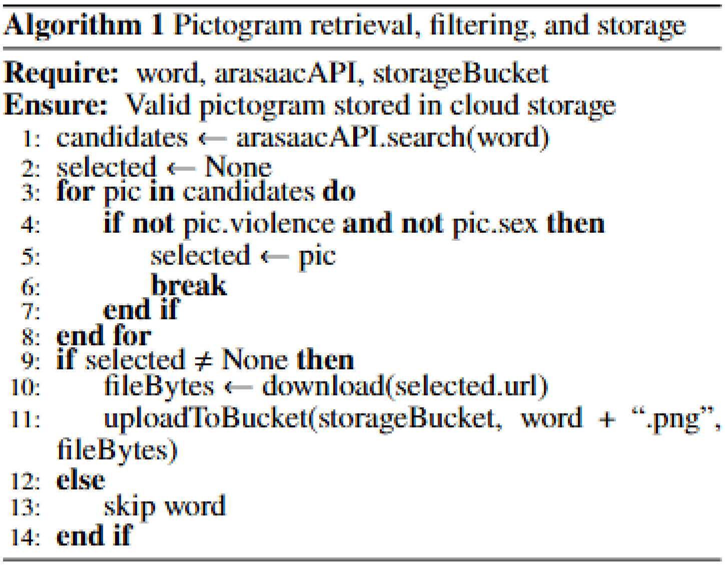
From an infrastructural standpoint, the back-end was containerized using Docker and deployed on Google Cloud Run. This environment was selected for its automatic scaling and cost-efficiency: the server remains inactive during idle periods and is automatically reactivated upon incoming requests, significantly reducing operational overhead during experimentation.

User authentication and data management are handled through API endpoints that interface with a PostgreSQL database hosted on Neon. During user registration, the back-end hashes passwords and persists user records in the database. At login, the system verifies the submitted credentials and, when successful, issues a JWT token, which is then required to access protected routes. This mechanism ensures that each user can access only their own albums and generated boards, which is essential for preserving data privacy, particularly due to the involvement of images of children.

A central responsibility of the back-end is the execution of the communication board generation pipeline. Upon receiving an image from the front-end, the server forwards the content to the artificial intelligence module responsible for producing short contextual stories. After receiving the generated stories, the back-end extracts relevant words, creates a new album entry in the database, and initializes a private folder in Google Cloud Storage. The original image is stored in this folder along with the future pictograms and metadata.

For each extracted word, the back-end queries the ARASAAC API, which may return multiple pictogram candidates. Because ARASAAC includes pictograms intended for different age groups and contexts, each candidate is evaluated according to its metadata. Only pictograms that do not contain sensitive content—such as violence or sexual representations—are stored. This filtering logic is essential for ensuring suitability for non-verbal children aged 5 to 8 years. The selection and storage procedure is summarized in Algorithm (Figure 4)

Figure 4. Pictogram selection.



Algorithm 1 Pictogram retrieval, filtering, and storage

```
Require: word, arasaacAPI, storageBucket
Ensure: Valid pictogram stored in cloud storage
candidates ← arasaacAPI.search(word)
selected ← None
for pic in candidates do
    if not pic.violence and not pic.sex then
        selected ← pic
        break
    end if
end for
if selected ≠ None then
    fileBytes ← download(selected.url)
    uploadToBucket(storageBucket, word + ".png",
fileBytes)
else
    skip word
end if
```

Finally, to deliver the generated communication boards to the front-end, the back-end issues **signed URLs** with temporary validity, granting controlled access to otherwise private resources. Once a URL expires, the client must request a new signed link, thus ensuring the confidentiality of stored images and metadata.

2.5.9 AI Module

The AI module is responsible for generating short, context-specific stories from the input image, which are subsequently used to retrieve the corresponding pictograms. This component relies on a multimodal large language model (LLM), chosen for its ability to jointly interpret visual and textual information. Unlike traditional computer vision pipelines that depend on large domain-specific datasets, multimodal LLMs benefit from broad and diverse pre-training corpora, allowing them to generalize across heterogeneous inputs such as real photographs, children's drawings, and cartoon characters.

Instruction Prompt Used in the System.

To standardize the linguistic output and ensure compatibility with AAC/PECS pictograms, the story-generation component used a fixed instruction prompt. The prompt constrained the model to produce short, simple, developmentally appropriate narratives in Portuguese. The exact English translation of the instruction is presented below:

“Generate two short educational stories in Portuguese, each with up to three sentences, for a 6-year-old autistic child. Use only simple words compatible with PECS/AAC pictograms. Begin each story with an action verb. If there is a character in the image, refer to them as he, she, or it. If there is no character, create the story as if a 6-year-old child were experiencing the scene. Do not use names, do not use punctuation, and separate the stories using ####. Respond only with the raw text.”

This fixed instruction was applied to every image submitted to the system and determined the structure of the textual output used for subsequent pictogram retrieval.

Two models were evaluated in this study: Gemini 2.0 Flash (Google) and the open-source Liu et al. LLaVA-Next. Their suitability for integration into the system was assessed through two small but focused diagnostic tests, one targeting character recognition and another targeting story generation under AAC constraints.

In the character-recognition stage, each model was presented with three images containing well-known characters from different generations of children’s media. Across these images, the set of characters included Naruto, Sakura, Sasuke, Superman, Sonic, Luna from the cartoon “Earth to Luna!" and Baby Shark. For every image, both models received the same prompt:

“What is this cartoon and which characters are in the picture?"

The goal was to verify whether the model could correctly identify the cartoon and list the characters present in the scene. Both models correctly recognized established characters such as Naruto, Sakura, Sasuke and Superman. However, LLaVA-Next showed consistent failures when dealing with more recent or preschool-oriented content, such as Baby Shark and Luna, often producing unrelated descriptions or omitting the characters altogether. Gemini, in contrast, identified all characters correctly in every tested image, suggesting a broader and more up-to-date multimodal training base.

In the story-generation stage, each model was again evaluated with three images drawn from the same set of characters, now focusing on its ability to produce AAC-compatible narratives. For every image, the models received the following prompt:

“Generate a short story for a 6-year-old autistic child about the character, with a maximum of 2 sentences."

The stories were assessed according to three practical criteria: compliance with the two-sentence limit, simplicity of vocabulary compatible with pictogram-based communication, and coherence with the visual context of the image. LLaVA-Next frequently violated the requested constraints, producing long descriptive paragraphs or, in some cases, failing to generate an appropriate story for specific images (for example, when analyzing scenes from ``Earth to Luna!"). Gemini consistently produced short, coherent and developmentally appropriate narratives, respecting the length constraint and maintaining a simple, concrete vocabulary suitable for use in AAC boards.

Based on recognition accuracy, linguistic adequacy, and practical deployment feasibility in the proposed pipeline, Gemini 2.0 Flash was selected as the primary model for the system. Although LLaVA-Next offers the advantage of being fully open-source and customizable, its baseline performance was less reliable, particularly for contemporary children’s content and for the strict narrative constraints required in this application. This limitation could be mitigated in future work through targeted fine-tuning, since the model can be retrained using datasets specifically tailored to cartoons, characters and AAC-oriented contexts. However, maintaining an open-source multimodal model in production would require continuous GPU hosting, significantly increasing operational costs. For the purposes of this study, Gemini provided a more stable and cost-efficient alternative for integration into the end-to-end communication board generation system.

2.6 Results

The results are presented in two parts: (1) the quantitative evaluation completed by AAC professionals and (2) the qualitative observations collected during the child interaction sessions.

2.6.1 Quantitative Results

Table 2 and Table 3 present the mean scores assigned to each questionnaire item and to each evaluation category. The highest individual value was obtained in "Long-term communication benefits" (4.0), followed by "System response time", "Increased child interest" and "Overall therapeutic usefulness" (all 3.8). The lowest averages were observed for "Story clarity and adequacy" (2.6).

Table 2. Mean scores for each questionnaire item (Likert scale 1–4; n=5).

Question	**Mean Score**
Q1: Pictogram quality	2.8
Q2: Story clarity and adequacy	2.6
Q3: Interface ease of use	3.4
Q4: Caption clarity for clinical use	3.0
Q5: System response time	3.8
Q6: Increased child interest	3.8
Q7: Facilitates communication	3.75
Q8: Reduced frustration	3.0
Q9: Complements other AAC strategies	3.75
Q10: Overall therapeutic usefulness	3.8

Q11: Long-term communication benefits	4.0

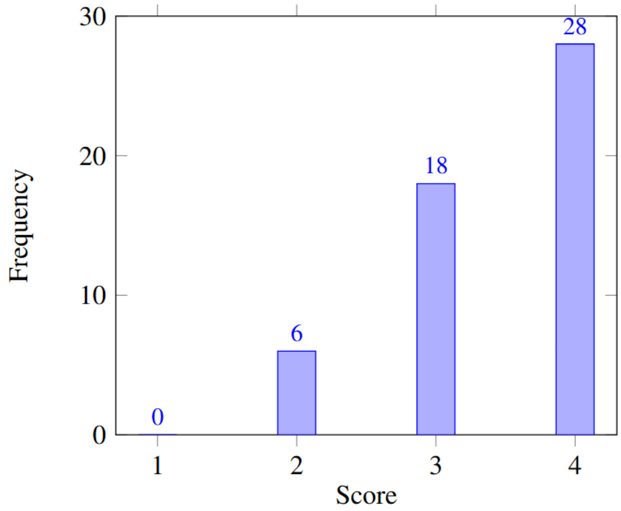
Category-level averages followed a similar pattern. As shown in table 3, the highest grouped score corresponded to "Integration and Professional Applicability" (3.9), while "Pictogram and Story Quality" presented the lowest mean (2.7).

Table 3. Mean scores grouped by evaluation category.

Category	**Mean Score**
Pictogram and Story Quality	2.7
Usability and System Performance	3.6
Therapeutic Impact	3.4
Integration and Professional Applicability	3.9

To complement these averages, Figure 4 shows the distribution of all ratings across the Likert scale. A total of 52 individual ratings were recorded. As shown in Figure 4, the distribution was concentrated in the upper half of the scale: 28 ratings (53.8%) corresponded to the maximum score of 4, and 18 ratings (34.6%) were equal to 3. Together, scores of 3 and 4 accounted for 88.4% of all responses. 6 ratings (11.6%) received a score of 2, and no score of 1 was assigned.

Figure 5. Distribution of all scores assigned by AAC professionals (n = 5).



2.6.2 Qualitative Results

Three non-verbal autistic children (ages 5-8) participated in short, naturalistic interaction sessions with the platform. Across all cases, the AI consistently generated two short stories per image. The first story was generally accurate and reflected the visual content appropriately, while the second story frequently diverged from the scene and did not correspond to the uploaded image. This pattern was observed in every interaction session, regardless of the child, type of photo, or context.

Only periods of direct engagement with the interface were counted as “active use time.” Below, each child’s interaction profile is summarized, including the type of photos used and the behavior observed.

Child 1 (8–10 minutes of active interaction)

- **Images used:** 3 personal photographs.
- **Type of the photo:**
  - holding a water bottle;
  - drawing and painting an elephant in yellow;
  - children riding bumper cars.

- **Engagement**: strong auditory interest, repeatedly tapping pictograms especially “yellow.”
- **Behavior**: sustained focus, replayed sounds multiple times, explored the board spontaneously.
- **Outcome**: satisfactory interaction.

Child 2 (7–9 minutes of active interaction)

- **Images used:** 5 images.
- **Type of the photo:**
    - 2 drawings of children playing soccer;
    - 3 fictional characters: Hulk, Spider-Man, and Venom.
- **Engagement**: explored all generated boards; showed interest in hero-related vocabulary.
- **Behavior**: combined pictograms into simple sequences and corrected mismatches after verbal explanation.
- **Outcome**: satisfactory interaction.

Child 3 (less than 1 minute of active interaction)

- **Images used:** 4 static images.
- **Type of the photo:**
    - ball drawing;
    - ball-pit illustration;
    - alphabet chart;
    - vowel card.
- **Engagement**: no meaningful interaction with the pictogram board.
- **Behavior**: the child consistently directed attention to a mobile device playing videos; brief interaction occurred only through pressing keyboard keys to trigger sounds.
- **Outcome**: unsatisfactory interaction with the platform.

Notably, Child 3 demonstrated a clear preference for videos over images, which fully inhibited engagement with the generated communication boards.

2.7 Analysis or Discussion of Results

2.7.1 Interpretation of Quantitative Results

The quantitative findings indicate that 'Integration and Professional Applicability" was the highest-rated category (3.9), followed by "Therapeutic Impact" (3.4). This pattern suggests that AAC professionals perceive the system not as a replacement for existing communication boards, but as a complementary tool capable of increasing children’s motivation and engagement. The high scores in "Increased child interest", "Facilitates communication" and "Long-term communication benefits" reinforce this interpretation, indicating that the platform has potential to reduce frustration and support the development of communicative intent over time.

Regarding "Usability and System Performance", the platform received consistently positive evaluations (mean 3.6). Testers highlighted that the interface is intuitive and requires minimal instruction, and the response time was considered efficient, avoiding delays that could disrupt interaction. These characteristics are essential in AAC contexts, where user attention can be limited and interruptions may compromise engagement.

In contrast, the lowest scores were observed in "Pictogram and Story Quality" (2.7). Two specific issues explain this result. First, the pictogram retrieval occasionally introduced mismatches, such as returning the pictogram for "golfinho"" (dolphin) when the intended word was "gol" (goal). This occurred because the system occasionally selected the second-best pictogram from the ARASAAC database instead of the correct one, potentially generating semantic confusion for the child. Second, the model frequently produced an inaccurate second story for each image. The prompt did not explicitly constrain the model to remain within the visual content, leading the system to incorporate contextual assumptions about the child rather than describing the actual image. These limitations directly impacted the clarity and reliability of the generated narratives.

Despite these issues, the overall rating distribution shows that 88.4\% of all responses corresponded to scores of 3 or 4, demonstrating high satisfaction among testers and reinforcing the system’s practical usefulness. The results therefore indicate that the platform performs well in usability and therapeutic relevance, although targeted improvements are necessary in pictogram selection and story-generation accuracy.

2.7.2 Interpretation of Qualitative Findings

The qualitative results indicate that the platform is not designed for independent use by the child; instead, guided mediation by a caregiver or professional is essential to stimulate vocabulary learning and support communicative intent. For the first two children, the sessions were considered successful in terms of engagement. Although the interaction time was short to observe measurable improvements in communication, both children demonstrated sustained attention, explored multiple images, and interacted spontaneously with the pictogram board. In the second case, the child’s younger sister (approximately 4–6 years old) also showed interest by repeatedly selecting pictograms to hear the corresponding audio output. Although she was not an autistic or non-verbal child, her positive reaction suggests that the platform may be perceived as playful and intrinsically motivating by young users in general.

A relevant pattern observed for the first two children was the use of more than one image during the session. The willingness to test different photos indicates exploratory behavior and curiosity toward the platform, which is a positive indicator of engagement in AAC interventions. Both children maintained interaction for more than five minutes, which is noteworthy given the simplicity of the tool, which only displays pictograms and plays the corresponding audio without animations, gamified elements, or interactive feedback. Despite these minimal features, the children showed meaningful engagement: one child repeatedly listened to words that interested her, while the other created simple phrases related to fictional characters. These behaviors suggest that even limited interaction resources can be sufficient to sustain focus when the content is relevant to the child's personal interests.

In contrast, the third child did not engage meaningfully with the platform. She exhibited a strong pre-existing preference for videos and consistently redirected her attention away from the images. Her interaction time was extremely brief—less than one minute. Attempts to guide her toward the pictogram board resulted in visible frustration, preventing productive participation. This case highlights an important consideration in AAC contexts: children differ significantly in sensory profiles and motivational drivers. For this child, a video-based modality or tactile alternative might be more appropriate than static images on a touchscreen, as she interacted almost exclusively through keyboard presses without looking at the screen. This suggests that future versions of the system may benefit from multi-sensory features, such as textured interfaces or physical AAC devices, to support children with similar sensory preferences.

2.7.3 Cross-Analysis: Integrating Quantitative and Qualitative Evidence

Cross-analysis of the quantitative and qualitative results reveals several converging patterns regarding the system's usefulness and limitations. First, the high quantitative rating for "Increased child interest" is consistent with the qualitative observations: the first two children interacted with the platform for more than five minutes, explored multiple images, and treated the activity as a game. Although the system does not function as a conventional AAC board for expressing immediate needs (e.g., requesting water), it appears to complement existing AAC tools by encouraging children to explore new words and construct simple phrases. This exploratory engagement may be beneficial in therapeutic contexts, where motivation and sustained attention are essential components of language development.

The analysis also shows that "Story clarity" the lowest-rated quantitative item and directly affected the children's interaction patterns. The first story was consistently used to generate the pictogram board, while the second story was ignored by all participants due to its frequent semantic inconsistencies. The second child occasionally examined the second story only to identify additional vocabulary, but did not rely on it to interpret the image. This confirms that narrative inaccuracies reduce the practical usability of the tool and can limit the communicative value of the generated boards.

Another converging pattern concerns pictogram accuracy. The quantitative criticism of "Pictogram quality" aligns with qualitative observations: mismatched pictograms (e.g., "gol" generating "golfinho") caused confusion and were immediately avoided by the children. These mismatches hindered the learning process and demonstrated that semantic filtering is essential before deployment in real therapeutic environments.

Regarding "Usability and System Performance", the qualitative data strongly corroborate the high quantitative scores. Children had no difficulty understanding the interface: they simply pressed the pictograms and listened to the audio output without hesitation. The system also demonstrated a consistent response time of less than one minute for generating each board, allowing smooth continuation of the session and preventing loss of attention. However, the qualitative results also reveal that, despite good usability, caregiver mediation was always required to maintain engagement. This supports the interpretation that the tool is best suited for guided therapeutic sessions rather than autonomous use.

Finally, the quantitative item "Reduced frustration" could not be meaningfully evaluated. None of the engaged children showed signs of frustration during the activity; in contrast, the frustration observed in the third child was not caused by the system, but by the removal of a preferred video. Thus, the current study cannot conclusively determine whether the platform reduces communicative frustration, and longer multi-session studies will be needed to assess this dimension.

Together, these converging findings demonstrate that the system is promising as a complementary AAC resource that enhances engagement and supports vocabulary exploration, but requires improvements in story reliability and pictogram accuracy to maximize clinical applicability.

2.7.4 Relation to Prior Work

This study adopts a methodology that diverges from prior AAC research by employing a multimodal Large Language Model (LLM) to interpret diverse types of images, including cartoons, sketches, and fictional characters, generating short stories that serve as the basis for pictogram boards. This narrative-driven approach is particularly relevant for children, as it enables the system to adapt content to personally meaningful stimuli, an aspect that earlier AAC tools were not designed to support.

Vargas et al. emphasize the importance of collaborative and flexible tools capable of generating communication content directly from user-provided photos. The findings of the present study reinforce this perspective: generating pictogram boards from images proved to be motivating for children, supporting exploratory behavior and functioning as a complementary AAC resource rather than a replacement for existing systems.

Other AAC tools discussed in the literature, such as LIVOX, differ substantially from the approach proposed in this work. These systems are primarily designed for functional daily communication and can be used independently by the child once configured. In contrast, the proposed tool requires adult mediation and is not intended to function as a primary communication board. Instead, children interacted with the system in a more playful and exploratory manner, repeating words, responding to familiar vocabulary, and creating simple narrative elements. This suggests that the system may serve a distinct role within AAC practice: not as a standalone communication solution, but as a ludic, vocabulary-expanding resource that supports engagement and language exposure through interaction with personalized visual content.

3 Future Works

Several opportunities for improvement emerged from the findings of this study. First, narrative accuracy must be enhanced, particularly by refining the prompting strategy to prevent the model from generating story elements unrelated to the image. Future versions of the system could incorporate stricter visual grounding constraints or employ rule-based filters to suppress semantic drift in the second generated story.

Another important direction involves expanding the underlying dataset. Creating a curated image set including cartoons, drawings, fictional characters, and their corresponding captions would support more reliable performance, especially if used to fine-tune an open-source multimodal model. This could substantially improve story relevance and pictogram selection for child-focused content.

Additionally, the system may benefit from supporting video uploads rather than static images alone, allowing better alignment with children who show strong preference for dynamic visual stimuli. Improving the semantic validation of pictograms is also essential: future work could implement automated checks comparing generated story elements with ARASAAC captions to ensure that selected pictograms match the intended meaning.

Finally, the evaluation should be expanded through longer, continuous therapeutic use and a broader participant pool. The present sample was small, and longitudinal testing could reveal whether engagement persists over time and whether the tool contributes to observable communication gains in real-world AAC practice.

4 Conclusion

This exploratory study introduced a multimodal LLM-based approach for generating AAC boards and short educational stories from photos, drawings, and cartoon images. The results indicate that artificial intelligence can support AAC practice by producing personalized, interest-driven content that captures children's attention and promotes exploratory interaction. AAC professionals evaluated the platform positively, particularly highlighting its usability, responsiveness, and potential to complement existing communication strategies.

An important contribution of this work is the demonstration that multimodal LLMs are especially effective for interpreting drawings and cartoon characters—types of images that are highly relevant in child-centered AAC contexts. Unlike traditional computer-vision pipelines, which tend to struggle with stylized or non-photographic input, the model was able to extract meaningful visual elements from these images and generate vocabulary and stories that remained usable and engaging. This suggests that LLM-based approaches may expand AAC tools beyond real-world photographs, enabling richer and more personalized content for non-verbal children.

At the same time, the study also revealed limitations. Narrative accuracy was inconsistent, with the first story generally aligned with the image and the second frequently diverging from it. Occasional pictogram mismatches also reduced clarity when a pictogram did not correspond to the intended concept, children tended to ignore it, highlighting the importance of ensuring accurate and reliable symbol retrieval. Although none of the children used the platform independently, the first two engaged productively with adult mediation.

Overall, this work provides initial empirical evidence that multimodal LLMs can contribute meaningfully to child-centered AAC tools. Future research should focus on refining prompting methods, improving visual grounding and pictogram validation, expanding datasets to better support stylized imagery, and conducting longitudinal evaluations within therapeutic settings.

References

Article:

DI PAOLA, Ambra; MURARO, Serena; MARINELLI, Roberto; PILATO, Christian, **Foundation Models in Augmentative and Alternative Communication: Opportunities and Challenges**, *arXiv*, [S. l.], 2024, Available at: https://arxiv.org/abs/2401.08866, Accessed on: Dec. 18, 2025.

PARK, Chanjun; JANG, Yoonna; LEE, Seolhwa; SEO, Jaehyung; YANG, Kisu; LIM, Heuiseok**, PicTalky: Augmentative and Alternative Communication Software for Language Developmental Disabilities**, *arXiv*, [S. l.], 2022, Available at: https://arxiv.org/abs/2109.12941, Accessed on: Dec. 18, 2025.

CAMILO, Simoni; CRUZ, Fernanda Miranda da; CAETANO, Sheila C.; PERISSINOTO, Jacy; TAMANAHA, Ana Carina, **Pre-verbal and verbal pattern as predictors for the implementation of the Picture Exchange Communication System (PECS) in autistic children**, *Revista CEFAC*, [S. l.], v. 25, n. 6, p. e5823, 2023, Available at: https://doi.org/10.1590/1982-0216/20232565823, Accessed on: Dec. 18, 2025.

NEAMTU, Rodica; CAMARA, André; PEREIRA, Carlos; FERREIRA, Rafael, **Using Artificial Intelligence for Augmentative Alternative Communication for Children with Disabilities**, *Human-Computer Interaction – INTERACT 2019*, Cham, Springer, p. 234–243, 2019, Available at: https://doi.org/10.1007/978-3-030-29381-9_15, Accessed on: Dec. 18, 2025.

FONTANA DE VARGAS, Mauricio; DAI, Jiamin; MOFFATT, Karyn, **AAC with Automated Vocabulary from Photographs: Insights from School and Speech-Language Therapy Settings**, *Communications of the ACM*, New York, v. 68, n. 1, p. 89–96, 2025, Available at: https://doi.org/10.1145/3623505, Accessed on: Dec. 18, 2025.

FONTANA DE VARGAS, Mauricio; MOFFATT, Karyn, **Automated Generation of Storytelling Vocabulary from Photographs for Use in AAC**, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, Association for Computational Linguistics, p. 1353–1364, 2021, Available at: https://aclanthology.org/2021.acl-long.108/, Accessed on: Dec. 18, 2025.

LIU, Haotian; LI, Chunyuan; LI, Yuheng; LI, Bo; ZHANG, Yuanhan; SHEN, Sheng; LEE, Yong Jae, **LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge**, *arXiv*, [S. l.], 2024, Available at: https://llava-vl.github.io/blog/2024-01-30-llava-next/, Accessed on: Dec. 18, 2025.

CHARLOP-CHRISTY, M. H.; CARPENTER, M.; LE, L.; LEBLANC, L. A.; KELLET, K., **Using the Picture Exchange Communication System (PECS) with children with autism: Assessment of PECS acquisition, speech, social-communicative behavior, and problem behavior**, *Journal of Applied Behavior Analysis*, [S. l.], v. 35,

n. 3, p. 213–231, 2002, Available at: https://doi.org/10.1901/jaba.2002.35-213, Accessed on: Dec. 18, 2025.

WALLER, Annalu, **Telling tales: unlocking the potential of AAC technologies,** ***International Journal of Language & Communication Disorders***, [S. l.], v. 54, n. 2, p. 159–169, 2019, Available at: https://doi.org/10.1111/1460-6984.12449, Accessed on: Dec. 18, 2025.