

Evaluating Therapist-in-the-Loop NLP in Comprehension Rehabilitation for Children with SEND

Martina Galletti

Sony Computer Sciences Laboratories - Paris, France
martina.galletti@sony.com

Caterina Marchesi

Centro Ricerca e Cura di Roma, Italy

Eleonora Pasqua

Centro Ricerca e Cura di Roma, Italy

Abstract

Personalized rehabilitation can improve educational and clinical outcomes for children with special educational needs and disabilities (SEND), but creating individualized reading exercises, comprehension questions, lexical supports, and progress reports remains time-consuming for clinicians. We present ARTIS, a therapist-in-the-loop and natural-language-processing (NLP)-assisted platform that supports reading-comprehension rehabilitation while preserving expert oversight, and evaluate its deployment in a real clinical setting. We report a ten-session intervention involving ten children aged 9–11 years at a speech and language rehabilitation centre. Across the intervention, platform comprehension accuracy remained relatively stable, while pre/post analyses showed stronger gains in vocabulary knowledge and lexical competence than in standardized reading-comprehension scores. Changes in platform accuracy were positively associated with changes in standardized comprehension outcomes, although these exploratory results should be interpreted cautiously. Overall, the study demonstrates the feasibility of integrating therapist-supervised NLP into a real rehabilitation workflow and highlights its potential to facilitate personalized material preparation while maintaining clinician oversight.

1 Introduction

Large language models and other NLP technologies create new opportunities for accessible education and speech and language rehabilitation particularly when individualized materials must be produced repeatedly for learners with diverse needs. However, positive impact in high-stakes domains cannot be established by generation quality alone. NLP systems intended for children, education, and clinical rehabilitation must be evaluated in realistic settings, designed around professional oversight, and assessed in terms of whether they support meaningful human outcomes.

Reading comprehension rehabilitation for children with SEND is one such setting. Children with SEND often require individualized texts, questions, lexical explanations, visual supports, and repeated practice. Preparing such materials manually places a substantial burden on therapists and limits the scalability of personalized intervention (Balaskas et al., 2024). NLP can assist this process by generating candidate questions and lexical activities, but unsafe or inappropriate generated content would be unacceptable in clinical practice and dangerously counter-productive. For this reason, ARTIS is designed as a therapist-in-the-loop system: NLP assists material creation and progress monitoring, while speech and language therapists retain responsibility for validation, adaptation, and therapeutic use.

This paper reports a clinical deployment of ARTIS in collaboration with Centro Ricerca e Cura di Roma, a speech and language therapy center located in Rome and affiliated to the Italian national health system. ARTIS integrates therapist-selected texts, GPT-generated comprehension questions, lexical support, pictograms, semantic-network exercises, and progress reports into a structured rehabilitation workflow. The study evaluates whether this NLP-assisted workflow is feasible in routine rehabilitation and whether platform-derived measures align with standardized clinical outcomes.

Our contribution is not a new language model or generation algorithm: the core ARTIS architecture and NLP components were introduced in earlier work (Galletti et al., 2023; Galletti and Cesaroni, 2025; Galletti, 2026; Jhilar and Galletti, 2026; Jhilar et al., 2026). The present work does not introduce a major redesign of the underlying ARTIS architecture; its novelty lies in the longitudinal clinical deployment and evaluation of the existing therapist-in-the-loop workflow. We evaluate this existing therapist-in-the-loop workflow through a ten-session clinical deployment with chil-

dren with SEND. We contribute longitudinal interaction data, comparisons with standardized clinical assessments, and practical evidence and limitations concerning therapist-supervised NLP deployment in rehabilitation.

2 Background and Related Work

2.1 AI for Reading Support and Rehabilitation

Recent NLP and AI systems have been explored to support reading through contextual definitions, semantic highlighting, automatic summarization, and adaptive text simplification. Examples include Semantic Reader (Lo et al., 2023), SciReader (Head et al., 2021), and gaze-driven text simplification approaches (Higasa et al., 2023). While these systems improve reading accessibility, they primarily function as assistive technologies rather than rehabilitation tools, and they do not provide structured therapeutic activities or clinician-supervised interventions. Interactive educational systems such as 3D Readers¹ (Johnson-Glenberg, 2007) and CACSR (Kim et al., 2006; Stetter and Hughes, 2011) have also demonstrated the benefits of strategy-based reading instruction through question generation, visual supports, and personalized feedback. However, these approaches generally do not exploit recent NLP techniques for automatic content generation or target clinical rehabilitation. Several digital and tele-rehabilitation platforms have been proposed for children with dyslexia and related reading difficulties (Wang et al., 2022; Zingoni et al., 2021; Barua et al., 2022; Rajapakse et al., 2018). These systems mainly focus on decoding and word-level skills, with limited support for higher-level comprehension processes such as inference generation and semantic integration. In Italy, RIDInet² provides clinically oriented online rehabilitation exercises, but does not employ NLP-based content generation or adaptive language technologies. ARTIS addresses this gap by combining NLP-assisted generation of comprehension questions, lexical supports, and multimodal scaffolding with therapist supervision.

2.2 Reading Comprehension Difficulties

Reading comprehension difficulties constitute a heterogeneous clinical profile that is partially independent of decoding ability (Oakhill et al., 2003;

Stothard and Hulme, 1995, 1996). Children may experience difficulties integrating information, generating inferences, or constructing coherent mental representations of a text despite adequate decoding skills (American Psychiatric Association, 2013; Nation and Snowling, 2000). Conversely, dyslexia primarily affects decoding and reading fluency, although comprehension difficulties may co-occur in some individuals (Snowling, 2000; Nation, 2005). Beyond diagnosed disorders, many poor comprehenders also struggle with text understanding because of limited vocabulary, reduced reading experience, or broader cognitive factors (Golinkoff, 1976). These findings motivate rehabilitation approaches that support both lexical development and higher-level comprehension while adapting to heterogeneous learner profiles.

3 System Overview

ARTIS (Galletti et al., 2023; Galletti, 2026) provides separate therapist and child interfaces for therapist-supervised reading-comprehension rehabilitation. As illustrated in Figure 1, the therapist selects or uploads a reading passage, after which the platform generates candidate rehabilitation materials. Before assignment to a child, the therapist can inspect, edit, regenerate, or discard each generated item. AI-generated content is therefore never delivered directly to participants without clinical review. The NLP pipeline combines task-specific components (Galletti, 2026). Keywords are extracted using a KeyBERT-based model and linked to pictograms through the text-to-pictogram mapping pipeline (Jhilmil and Galletti, 2026). The source passage provides contextual example sentences, while gpt-3.5-turbo-0125 generates child-friendly lexical definitions, candidate comprehension questions, and semantically related terms and distractors for lexical exercises. For lexical definitions, the model is prompted to generate clear, age-appropriate explanations for children aged 6–11. For comprehension-question generation, gpt-3.5-turbo-0125 received the complete reading passage together with a fixed task-specific prompt requesting ten closed-ended comprehension questions. No task-specific model fine-tuning was performed, and all generated questions were clinically validated before use.

All materials used during the intervention were reviewed by speech and language therapists before use. ARTIS also includes coherence scoring,

¹<https://www.sbir.gov/node/334161>

²<https://www.anastasis.it/ridinet/>

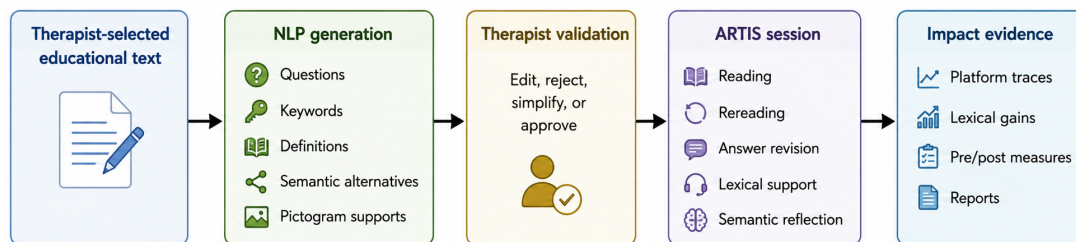


Figure 1: Therapist-in-the-loop workflow used by ARTIS.

sentence simplification, and concept-map generation, but these components were not part of the standardized intervention analyzed here. After therapist validation, the child first reads the selected text independently or with synchronized text-to-speech support. Comprehension questions are then presented through the interface. Following the initial response, the text is re-presented sentence by sentence with lexical and pictographic support, after which the participant may revise the response without being shown the correct answer. Participants subsequently complete lexical-definition and semantic-network activities. ARTIS records these interactions and generates longitudinal reports for therapist review.

4 Experimental Design

The study was conducted in collaboration with Centro Ricerca e Cura di Roma, a speech and language therapy center affiliated with the Italian national health system. The reading materials used during the intervention were drawn from educational resources designed for the second to fifth grades of primary school and routinely used in the clinical setting in this center. The ten texts were selected by the treating therapists from materials routinely used at the clinic and were matched to each participant's educational and clinical profile. Text selection was intended to maintain an appropriate and broadly comparable level of challenge across sessions rather than to impose a systematic increase in difficulty over time. Consequently, session number should not be interpreted as a formal progression from easier to harder texts. Nevertheless, because texts differed naturally in vocabulary, length, syntactic structure, and inferential demands, some between-session variation in difficulty is unavoidable and may contribute to fluctuations in weekly accuracy. For each selected text, ten closed-ended reading-comprehension questions were generated using gpt-3.5-turbo-0125 as detailed in

(Galletti, 2026) and manually validated by speech and language therapists before administration. The same therapist-supervised principle was applied to the other automatically generated materials used during the sessions: clinicians could inspect and modify the generated outputs before they became available to participants.

Ten children between 9 and 11 years of age, corresponding to grades three to five, participated in the intervention. Participation was voluntary and took place with informed consent from parents or legal guardians. Recruitment was carried out in collaboration with the clinical institution based on participants' diagnoses and therapeutic needs. Most participants presented comorbidity, with at least one diagnosis involving reading-comprehension impairment. Reported clinical profiles included mixed language disorders affecting expressive and receptive components, attention and behavioural-regulation difficulties, borderline intellectual functioning or mild intellectual disability, and specific learning disorders as it can be seen in Figure 2. The study followed a three-phase design: (1) Initial assessment, including standardized neuropsychological and linguistic evaluation together with a lexical test; (2) Experimental intervention, consisting of ten ARTIS sessions; (3) Final assessment, in which the same measures were repeated after the intervention.

Each participant first completed a standardized neuropsychological assessment to characterize cognitive and linguistic functioning. The Wechsler Intelligence Scale for Children (WISC) was used to assess general psychological functioning and neuropsychopathological disorders, including verbal comprehension, visual-perceptual reasoning, working memory, and processing speed. Reading and language skills were assessed with two standardized tools: (1) MT-3 Clinica, including measures of reading accuracy, reading speed, and reading comprehension; (2) BVL 4–12, covering multiple

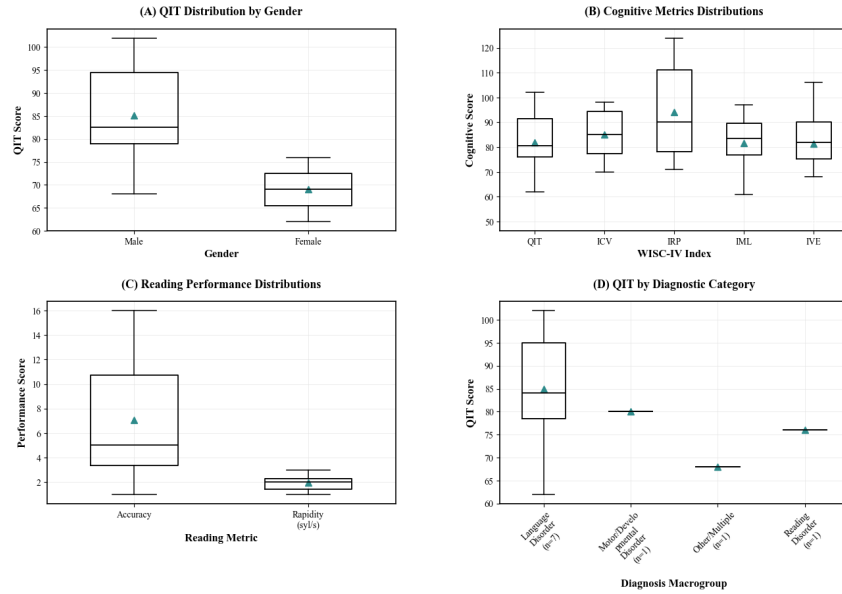


Figure 2: Boxplots display medians (horizontal line) and means (triangle). *QIT* stands for Total Intelligence Quotient, *ICV* for Verbal Comprehension Index, *IRP* for Visual-Perceptual Reasoning Index, *IML* for Working Memory Index, and *IVE* for Processing Speed Index.

(a) MT-3 Text Comprehension

Participant	Text 1 (Pre)	Percentile (Pre)	Text 2 (Pre)	Percentile (Pre)	Total (Pre)
User1	4/12	5°	2/12	<5°	6/24 <5°
User2	5/12	10°	6/12	15°–20°	11/24 15°
User3	3/12	5°	5/12	10°–20°	11/24 5°
User4	5/12	15°–20°	5/12	30°	10/24 15°
User5	5/12	15°–20°	4/12	10°–20°	9/24 10°
User6	8/12	50°–60°	4/12	10°–15°	12/24 20°
User7	6/12	15°–20°	4/12	10°–15°	10/24 10°
User8	4/12	5°	5/12	15°–20°	9/24 10°
User9	3/12	<5°	3/12	<5°	6/24 <5°
User10	5/12	15°–20°	3/12	5°	8/24 <5°

(b) BVL — Expressive and Receptive Vocabulary

Participant	Expressive Vocabulary	Z-score	Receptive Vocabulary	Z-score
User1	63/67	0	N/A	N/A
User2	57/67	-1.5	N/A	N/A
User3	61/67	0	31/42	$-2 < z < -1.5$
User4	57/67	$-1 < z < 0$	34/42	$-1 < z < 0$
User5	61/67	0	33/42	-1
User6	61/67	0	34/42	0
User7	56/67	-1	32/42	-1
User8	49/67	-2	30/42	-2
User9	50/67	-2	28/42	-2
User10	60/67	$-1 < z < 0$	31/42	$z > 2$

Table 1: Baseline standardized reading and language assessment results. MT-3 scores report the number of correct answers for two reading-comprehension texts and the corresponding age/grade-referenced percentile ranges; lower percentiles indicate weaker performance relative to normative expectations. BVL scores report expressive and receptive vocabulary performance together with standardized z-score ranges, where values around 0 indicate performance close to the normative mean and negative values indicate below-average performance. N/A indicates that the measure was not administered or was unavailable.

aspects of language, with emphasis in this study on expressive and receptive vocabulary. In addition, therapists developed an internal lexical test based on keywords extracted from the texts used on the platform. Participants were asked whether they knew each word and, when applicable, to define it. Responses were scored on a 0–3 scale: 0 for

unknown words, 1 for incorrect definitions, 2 for partially correct or imprecise definitions, and 3 for correct definitions. The lexical test was administered before and after the ten-session intervention.

The intervention consisted of ten ARTIS sessions, each centred on a different text with ten associated comprehension questions, for a total of

ten texts. Sessions were conducted twice per week and, when necessary, remotely. Because sessions were conducted twice per week, each weekly value reported in Figure 4 aggregates performance across the session texts administered during that week. Thus, the weekly curve should not be interpreted as representing one fixed text per week. Each session followed the same standardized structure:

1. **Reading phase:** the participant read the text or listened to a synthesized version;
2. **Comprehension phase:** the participant answered closed-ended comprehension questions;
3. **Supported rereading:** the text was represented segmented into sentences and accompanied by pictograms as described in (Jhial and Galletti, 2026);
4. **Error review:** participants could revise earlier responses after rereading, without corrective feedback from the system;
5. **Lexical support:** participants accessed definitions, pictograms, and contextual sentences for selected keywords, then answered lexical questions as described in (Galletti, 2026);
6. **Semantic-network reflection:** for each keyword, participants identified semantically related alternatives among candidate options such as synonyms and distractors as described in (Galletti, 2026).

All ten participants completed the ten-session intervention. All sessions were conducted within a therapist-supervised rehabilitation setting. During in-person sessions, a speech and language therapist was present throughout the session, while remote sessions were conducted synchronously under therapist supervision. Therapists could provide procedural assistance, such as helping participants navigate the interface or access available supports, but did not provide correct answers to comprehension questions before participants responded or revised their answers. Accordingly, the results reported here reflect therapist-supervised use of ARTIS rather than fully independent use. No adverse events related to use of the platform were reported during the study. At the end of the intervention, therapists and participants could access PDF reports summarizing overall reading-comprehension

performance and temporal performance patterns across attempts.

To investigate the alignment between platform-based measures and clinical outcomes, we report a Spearman correlation comparing ARTIS accuracy metrics with MT-3 comprehension results. Owing to the exploratory nature of the study and the limited sample size, all statistical analyses should be interpreted as preliminary. We therefore emphasize descriptive statistics and effect sizes alongside inferential analyses, viewing the reported statistical results as evidence of alignment rather than definitive evidence of clinical efficacy.

5 Results

Baseline profile As shown in Table 1, participants showed heterogeneous baseline performance across MT-3 and BVL assessments, consistent with the clinical complexity of the sample. Pre-training MT-3 comprehension scores were generally low to moderate across the ten users, while BVL scores indicated variability in expressive and receptive vocabulary. The WISC-based overview likewise reflected broad variability in cognitive indices across the cohort.

In-platform reading-comprehension performance Across the intervention, ARTIS comprehension accuracy remained consistently high, ranging from 71.7% to 85.0% across attempts (Figure 3). Weekly averages showed a small positive descriptive trend across the intervention period (Figure 4). However, because each session involved a different text, performance reflects both participant-level change and natural variation in text and question difficulty. Although therapists selected materials to provide a broadly comparable and clinically appropriate level of challenge, the texts were not experimentally equated for lexical, syntactic, or inferential difficulty. The observed trend should therefore be interpreted as a longitudinal performance pattern rather than as a difficulty-controlled learning curve.

Within-session response revision We examined how responses changed when participants reconsidered comprehension questions after supported rereading (Figure 8). Among these within-session repeated responses, 70.1% remained correct, 11.2% changed from incorrect to correct, 12.6% remained incorrect, and 6.2% changed from correct to incorrect. These results describe immediate response sta-

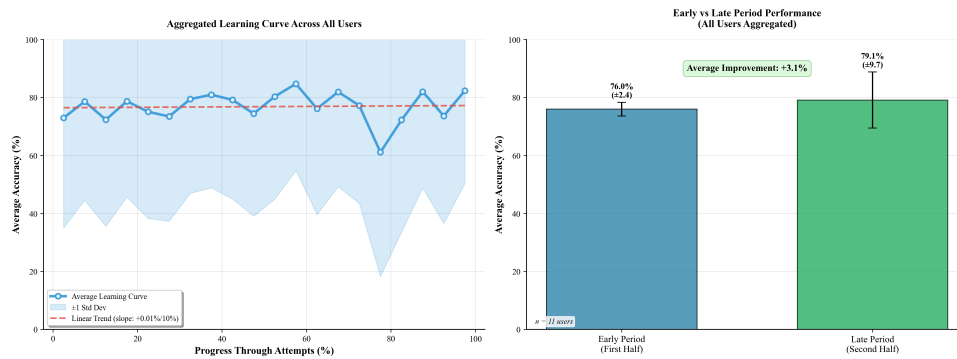


Figure 3: Longitudinal ARTIS comprehension performance and early-late comparison. The left panel shows aggregate accuracy across the ten intervention sessions, and the right panel compares mean accuracy during the first and final five sessions. Because different texts and questions were used across sessions, the figure describes longitudinal performance rather than a difficulty-controlled learning curve.

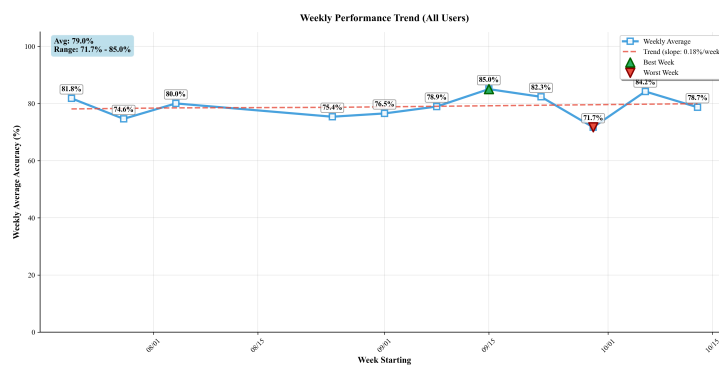


Figure 4: Weekly average accuracy across all users. Green and red markers identify the best and worst weeks, respectively.

bility and revision after access to additional reading support. Because the questions were reconsidered within the same session rather than after a delayed interval, this analysis should not be interpreted as evidence of long-term retention or learning across the intervention.

Pre/post assessment outcomes Finally, pre/post outcomes differed by domain. Standardized comprehension scores were variable, and several participants showed reduced post-treatment performance as we can see in Figures 5 and 6. Task difficulty or fatigue effects during post-treatment evaluation may have contributed to this pattern. By contrast, vocabulary-related outcomes were more consistently positive as we see in Figure 5-B. Nearly all participants showed either stable or slightly improved post-treatment performance in vocabulary expression and recognition. The internally developed lexical test reinforced this trend: all users improved from pre- to post-test on normalized lexical scores. The largest gains were reported for User 1 (approximately 58%), followed by User 10 (approximately 29%) and User 2 (approximately 23%), with most remaining participants showing

increases in the 10%–22% range. The proportion of known words also increased after the intervention for all users. In the most marked case, User 1 increased from approximately 41% known words at pre-test to approximately 91% at post-test. Users with lower baseline values also showed visible gains, indicating that the lexical component of the intervention was beneficial even for participants starting from weaker initial performance.

Association between platform results and clinical outcomes To investigate whether performance recorded by ARTIS aligned with standardized clinical measures, we compared platform accuracy with MT-3 reading-comprehension scores. We defined *early platform accuracy* as the mean ARTIS accuracy across the first five intervention sessions and *late platform accuracy* as the mean accuracy across the final five sessions. These measures should not be interpreted as pre- and post-treatment assessments: baseline MT-3 was administered before the intervention, whereas early ARTIS accuracy was measured during the first half of the intervention. Figure 7 summarizes three complementary analyses. Panel A compares baseline MT-3 compre-

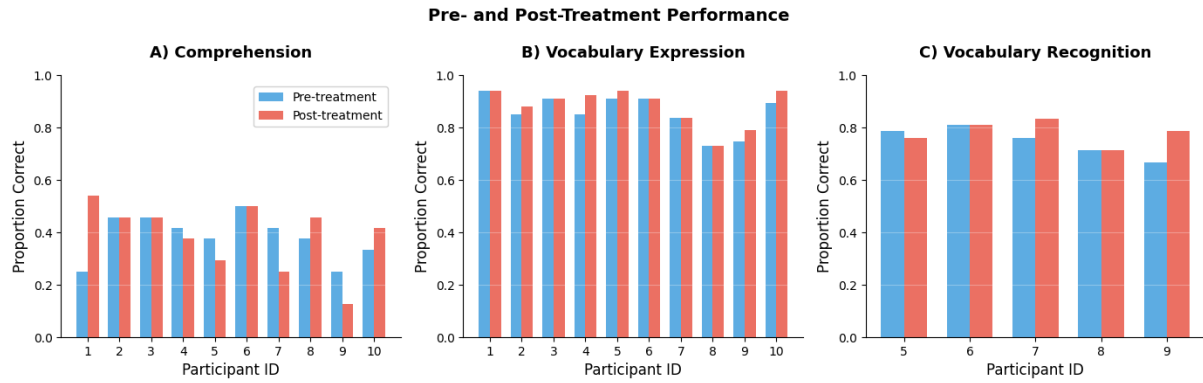


Figure 5: Pre- and post-treatment performance across comprehension, vocabulary expression, and recognition tasks. Subfigure C includes only participants with complete paired receptive-vocabulary measurements ($n = 5$).

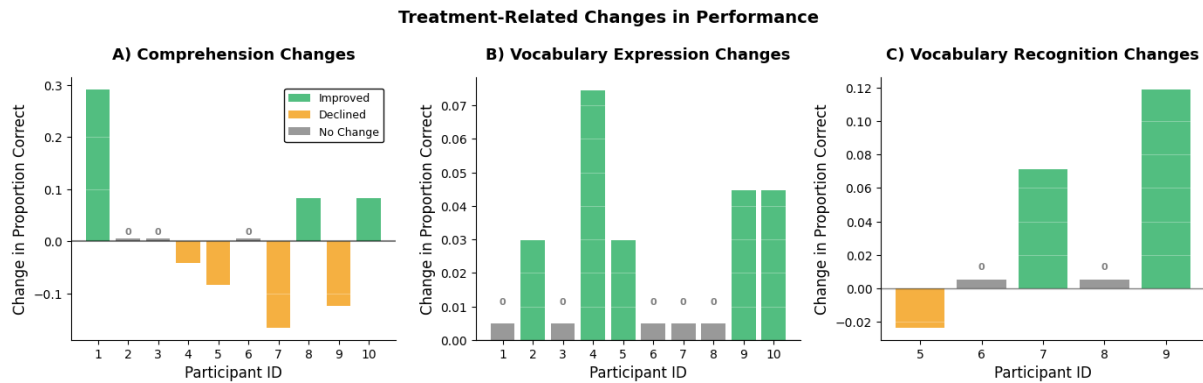


Figure 6: Individual treatment-related changes in each domain. Subfigure C includes only participants with complete paired receptive-vocabulary measurements ($n = 5$).

hension scores with early ARTIS accuracy. This association was weak and non-significant ($\rho = 0.092$, $p = 0.400$). Panel B compares post-intervention MT-3 scores with late ARTIS accuracy, showing a moderate but non-significant positive association ($\rho = 0.350$, $p = 0.161$). Panel C compares the change in MT-3 comprehension between pre- and post-assessment with the corresponding change in ARTIS accuracy from the early to the late intervention period. This analysis showed a positive association ($\rho = 0.656$, $p = 0.020$). Given the small and clinically heterogeneous sample, these correlations should be interpreted as exploratory. In particular, the improvement-based association provides preliminary evidence that changes captured by ARTIS may align with changes in standardized comprehension performance, but it does not establish ARTIS accuracy as a validated clinical outcome measure. Spearman's rank correlation was used because of the limited sample size and because the analysis did not assume normally distributed scores. The regression lines in Figure 7 are included only as visual aids and are not the basis

of the inferential analysis. We focused the correlation analysis on MT-3 comprehension because it measures text-level reading comprehension, which is the construct most directly corresponding to ARTIS comprehension accuracy. BVL measures were instead analyzed as pre/post language outcomes because they primarily assess expressive and receptive lexical abilities rather than the same task-level comprehension construct.

6 Lessons for Positive-Impact NLP Deployment

Clinical usefulness depends on expert control.

Human oversight is essential for high-stakes NLP applications. The platform was deliberately deployed as a decision-support tool rather than as an autonomous content provider. Because every generated item was reviewed before clinical use, inappropriate or insufficiently tailored outputs could be corrected or discarded before reaching participants. However, the present study did not systematically record the frequency or type of therapist modifications. We therefore cannot quantify how

often generated materials required correction or determine the standalone appropriateness of the underlying language-model outputs. This design demonstrates the feasibility of integrating generation into a clinician-supervised workflow, but it should not be interpreted as evidence that unreviewed model outputs would have been suitable for direct use with children. Rather than replacing clinical expertise, the NLP system was designed to assist material preparation while preserving human responsibility for therapeutic decisions.

NLP may facilitate material-preparation workflows. Informal feedback was obtained from four speech and language therapists involved in the deployment across meetings and suggested that automated generation and annotation facilitated material preparation. However, feedback was not collected through a structured qualitative instrument, and preparation time, editing workload, acceptance rates, and correction frequency were not systematically measured. The present study therefore establishes deployment feasibility rather than a quantified reduction in therapist workload; these measures should be included in future evaluations.

Platform metrics should be interpreted with clinical context. ARTIS produced interpretable traces of performance, including accuracy over time and within-session response-revision patterns. These traces were useful because they could be compared with standardized assessments and interpreted by therapists. However, the small sample and heterogeneous profiles mean that platform accuracy should not be treated as a standalone clinical outcome.

Lexical support may be an especially promising target. The clearest gains appeared in lexical expression, vocabulary recognition, and the internal lexical test. This suggests that NLP-generated definitions, contextual examples, pictograms, and semantic alternatives may be particularly useful in rehabilitation settings where vocabulary knowledge is a barrier to comprehension.

7 Discussion

This study contributes evidence that therapist-supervised NLP can be integrated into real rehabilitation workflows while maintaining expert oversight. Rather than evaluating language models solely through intrinsic generation quality, we assess their role within a complete human-centered

intervention in which clinicians retain responsibility for validating generated materials and interpreting learner progress. From a positive-impact NLP perspective, the primary contribution of this work is demonstrating the feasibility of responsibly deploying generative NLP technologies in a high-stakes educational setting. The observed improvements were more consistent for vocabulary-related measures than for standardized reading-comprehension assessments. Although these findings should not be interpreted as evidence of clinical efficacy, they suggest that NLP-assisted lexical support—including automatically generated definitions, contextual examples, and semantic activities—may represent a particularly promising direction for future rehabilitation research. The positive association between changes in ARTIS accuracy and changes in MT-3 comprehension is also noteworthy, but should be considered exploratory. It suggests that platform-derived performance may align with changes in standardized comprehension scores; however, the small and clinically heterogeneous sample does not allow us to establish the validity of ARTIS accuracy as a measure of cognitive change.

An important open question is whether the NLP-assisted workflow provides advantages over conventional therapist-prepared rehabilitation, including measurable reductions in material-preparation effort. The present study was not designed as a comparative effectiveness trial and therefore cannot establish that children achieve better outcomes with NLP-assisted materials than with standard therapist-created materials. The results instead demonstrate that personalized, therapist-validated materials can be delivered through the platform while preserving clinical oversight and generating longitudinal interaction data. Establishing the incremental benefit of NLP will require a future controlled comparison between NLP-assisted rehabilitation and an otherwise comparable therapist-led condition without automatic generation. Such a study should assess both child-level outcomes and therapist-level measures, including preparation time, editing effort, acceptance rates, perceived usefulness, and satisfaction.

8 Conclusion

We presented a therapist-in-the-loop NLP platform supporting personalized reading-comprehension rehabilitation for children with SEND and reported its deployment in a real clinical setting. The study

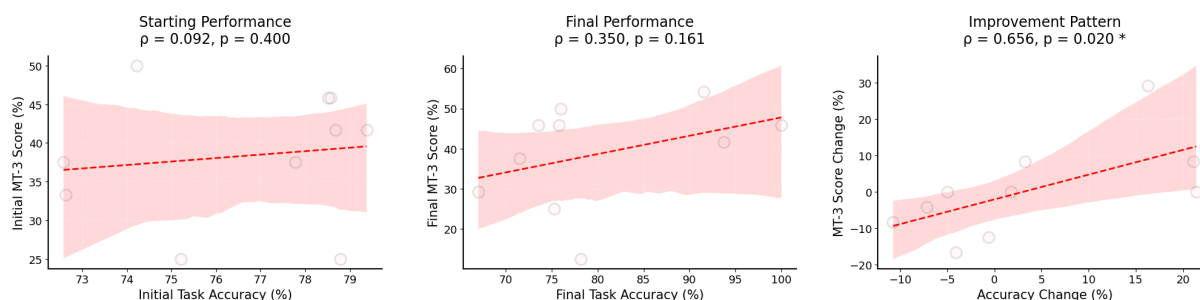


Figure 7: Associations between standardized MT-3 comprehension scores and platform accuracy. Panel A compares baseline MT-3 scores obtained before the intervention with mean platform accuracy during the first five sessions (“early” accuracy). Panel B compares post-intervention MT-3 scores with mean platform accuracy during the final five sessions (“late” accuracy). Panel C compares the pre/post change in MT-3 comprehension with the change in platform accuracy between the early and late intervention periods. Reported coefficients are Spearman rank correlations. Dashed regression lines are provided for visualization only.

provides preliminary evidence for the practical feasibility of integrating LLM-assisted content generation into routine rehabilitation while preserving clinician oversight through systematic human validation. Although the clinical evaluation is exploratory and limited by a small, heterogeneous cohort without a control group, the findings suggest that therapist-supervised NLP may facilitate the preparation of personalized rehabilitation materials and generate interpretable interaction data that complement standardized clinical assessments. Beyond the specific application, this work illustrates an emerging paradigm for positive-impact NLP in which language technologies augment professional expertise instead of replacing it.

Limitations

The study should be interpreted as preliminary. The sample comprised only ten participants and included heterogeneous clinical profiles and comorbidities, limiting the generalizability of the results and preventing meaningful analyses across diagnostic subgroups. Some decreases observed in post-intervention comprehension scores may also reflect assessment-related factors such as fatigue or task difficulty.

A second limitation is the absence of a concurrent control group. Consequently, changes observed during or after the intervention cannot be unequivocally attributed to ARTIS. Standardized assessments with established normative data provide a useful reference for interpreting participant performance (Capodieci et al., 2020), but such normative benchmarks cannot substitute for a matched or randomized control condition. Future studies

should therefore include larger samples, more balanced clinical subgroups, and an appropriate comparison condition to strengthen causal inference.

A further limitation concerns variation in the materials administered across sessions. Participants worked on different texts throughout the intervention, selected by therapists according to educational level and clinical needs. Although the materials were intended to provide a broadly comparable level of challenge, their linguistic and inferential difficulty was not experimentally equated. Session-to-session variation in platform accuracy may therefore partly reflect differences between texts and questions in addition to changes in participant performance.

Finally, the present study evaluates the complete therapist-in-the-loop workflow rather than the intrinsic quality of the NLP components in isolation. All generated materials were reviewed and could be modified before clinical use, but therapist edits, rejection rates, regeneration frequency, and preparation time were not systematically logged. We therefore cannot quantify either the standalone appropriateness of the generated outputs or the magnitude of the reduction in therapist workload. Future work should measure generation acceptance rates, editing time, failure types, and preparation effort, as well as compare alternative language models under equivalent clinical conditions.

Ethics Statement

Participation was voluntary and took place with informed consent from parents or legal guardians. The study was conducted within the framework of a formal collaboration agreement between Sony

Computer Science Laboratories - Paris and Centro Ricerca e Cura di Roma, which governed the respective responsibilities of the participating organizations for the clinical deployment of ARTIS and the management of study-related activities. The intervention was carried out within the clinic's rehabilitation setting and under the responsibility and supervision of qualified speech and language therapists. ARTIS was used exclusively as a therapist-supervised support tool and did not autonomously diagnose participants, determine eligibility, or make treatment decisions. AI-generated rehabilitation materials were reviewed by clinicians before being made available to children. The study therefore evaluated the use of NLP within an existing therapist-led rehabilitation workflow rather than replacing clinical judgment with automated decision making.

References

- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Publishing, Arlington, VA.
- Andreas Balaskas, Stephen M. Schueller, Kevin Doherty, Anna L. Cox, and Gavin Doherty. 2024. *Designing personalized mental health interventions for anxiety: Cbt therapists' perspective*. *International Journal of Human-Computer Studies*, 190:103319.
- Prabal Datta Barua, Jahmunah Vicnesh, Raj Gururajan, Shu Lih Oh, Elizabeth Palmer, Muhammad Mokhzaini Azizan, Nahrizul Adib Kadri, and U Rajendra Acharya. 2022. *Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review*. *International Journal of Environmental Research and Public Health*, 19(3):1192.
- Agnese Capodiecici, Cesare Cornoldi, Elizabeth Doerr, Laura Bertolo, and Barbara Carretti. 2020. *The use of new technologies for improving reading comprehension*. *Frontiers in Psychology*, 11:751.
- Martina Galletti. 2026. *A Computational Framework for AI-Driven Reading Comprehension Rehabilitation for Neurodiverse Learners*. Ph.D. thesis, Università degli Studi di Roma La Sapienza.
- Martina Galletti and Valeria Cesaroni. 2025. *From end-users to co-designers: Lessons from teachers*. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 505–516. Association for Computational Linguistics.
- Martina Galletti, Eleonora Pasqua, Francesca Bianchi, Manuela Calanca, Francesca Padovani, Daniele Nardi, and Donatella Tomaiuolo. 2023. *A reading comprehension interface for students with learning disorders*. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 282–287.
- Roberta Michnick Golinkoff. 1976. *A comparison of reading comprehension processes in good and poor comprehenders*. *Reading Research Quarterly*, 11(4):623–659.
- Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. *Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Taichi Higasa, Keitaro Tanaka, Qi Feng, and Shigeo Morishima. 2023. *Gaze-driven sentence simplification for language learners: Enhancing comprehension and readability*. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 292–296. Association for Computing Machinery.
- Soufiane Jhilal and Martina Galletti. 2026. *Robust multilingual text-to-pictogram mapping for scalable reading rehabilitation*. *arXiv preprint arXiv:2603.24536*.
- Soufiane Jhilal, Eleonora Pasqua, Caterina Marchesi, Riccardo Corradi, and Martina Galletti. 2026. *Tailoring ai-driven reading scaffolds to the distinct needs of neurodiverse learners*. In *International Conference on Artificial Intelligence in Education*, pages 505–520. Springer.
- Mina C. Johnson-Glenberg. 2007. *Web-based reading comprehension instruction: Three studies of 3d-readers*. In *Reading Comprehension Strategies: Theories, Interventions, and Technologies*, pages 293–324. Taylor and Francis.
- Ae-Hwa Kim, Sharon Vaughn, Janette K Klingner, Althea L Woodruff, Colleen Klein Reutebuch, and Kamiar Kouzekanani. 2006. *Improving the reading comprehension of middle school students with disabilities through computer-assisted collaborative strategic reading*. *Remedial and Special Education*, 27(4):235–249.
- Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, and 1 others. 2023. *The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces*. *arXiv preprint arXiv:2303.14334*.
- Kate Nation. 2005. *Children's reading comprehension difficulties*. In Margaret J. Snowling and Charles Hulme, editors, *The Science of Reading: A Handbook*, pages 248–265. Blackwell Publishing.
- Kate Nation and Margaret J. Snowling. 2000. *Factors influencing syntactic awareness skills in normal readers*

and poor comprehenders. *Applied Psycholinguistics*, 21(2):229–241.

Jane V. Oakhill, Kate Cain, and Peter E. Bryant. 2003. The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, 18(4):443–468.

Sampath Rajapakse, Dasuni Polwattage, Umesha Guruge, Isuru Jayathilaka, Tharindu Edirisinghe, and Samantha Thelijjagoda. 2018. Alexza: A mobile application for dyslexics utilizing artificial intelligence and machine learning concepts. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–6. IEEE.

Margaret J. Snowling. 2000. *Dyslexia: A Cognitive-Developmental Perspective*. Blackwell.

Maria Earman Stetter and Marie Tejero Hughes. 2011. Computer assisted instruction to promote comprehension in students with learning disabilities. *International Journal of Special Education*, 26(1):88–100.

Susan E Stothard and Charles Hulme. 1995. A comparison of phonological skills in children with reading comprehension difficulties and children with decoding difficulties. *Journal of Child Psychology and Psychiatry*, 36(3):399–408.

Susan E. Stothard and Charles Hulme. 1996. A comparison of reading comprehension and decoding difficulties in children. In Cesare Cornoldi and Jane V. Oakhill, editors, *Reading Comprehension Difficulties: Processes and Intervention*, pages 93–112. Lawrence Erlbaum Associates, Mahwah, NJ.

Min Wang, BalaAnand Muthu, and C. B. Sivaparthipan. 2022. Smart assistance to dyslexia students using artificial intelligence based augmentative alternative communication. *International Journal of Speech Technology*, 25:343–353.

Andrea Zingoni, Juri Taborri, Valentina Panetti, Simone Bonechi, Pilar Aparicio-Martínez, Sara Pinzi, and Giuseppe Calabrò. 2021. Investigating issues and needs of dyslexic students at university: Proof of concept of an artificial intelligence and virtual reality-based supporting platform and preliminary results. *Applied Sciences*, 11(10):4624.

A Supplementary Materials

This appendix presents example reports automatically generated by ARTIS for therapists, together with additional analyses of participant performance that complement the results presented in the main paper. These examples illustrate the type of feedback available to clinicians and provide further evidence of the platform’s ability to monitor learning progress throughout the intervention.

Figure 8 summarizes within-session changes in responses following supported rereading. Most initially correct responses remained correct (70.1%),

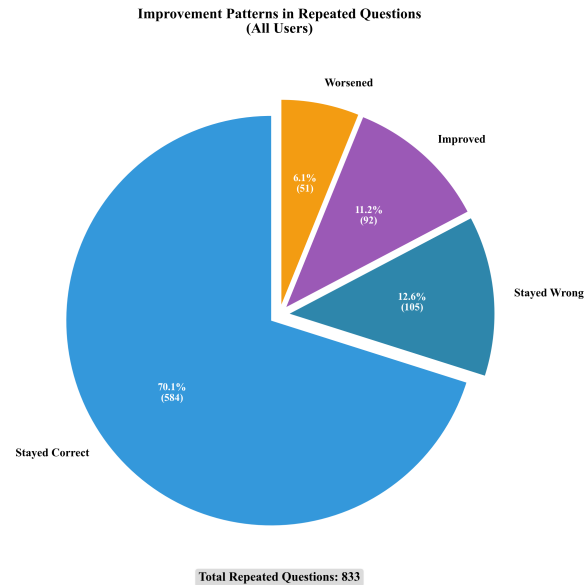


Figure 8: Within-session changes in comprehension responses following supported rereading, aggregated across participants. Responses are categorized as remaining correct, changing from incorrect to correct, remaining incorrect, or changing from correct to incorrect.

while 11.2% changed from incorrect to correct. Because questions were reconsidered within the same session, these outcomes reflect immediate response revision rather than delayed retention.

Figure 9 illustrates an example of the personalized report automatically generated after the intervention for an individual participant. The report summarizes overall reading-comprehension performance, highlights accuracy for each story, and visualizes the distribution of correct and incorrect responses. These summaries allow therapists to rapidly identify strengths, weaknesses, and texts requiring additional practice without manually reviewing every response.

Figure 10 provides a longitudinal view of the participant’s performance across rehabilitation sessions. The report compares early and late performance, summarizes weekly trends, and displays within-session response-revision patterns. These visualizations allow therapists to inspect longitudinal variation in platform performance, while recognizing that differences in text difficulty across sessions limit causal interpretation of changes over time.

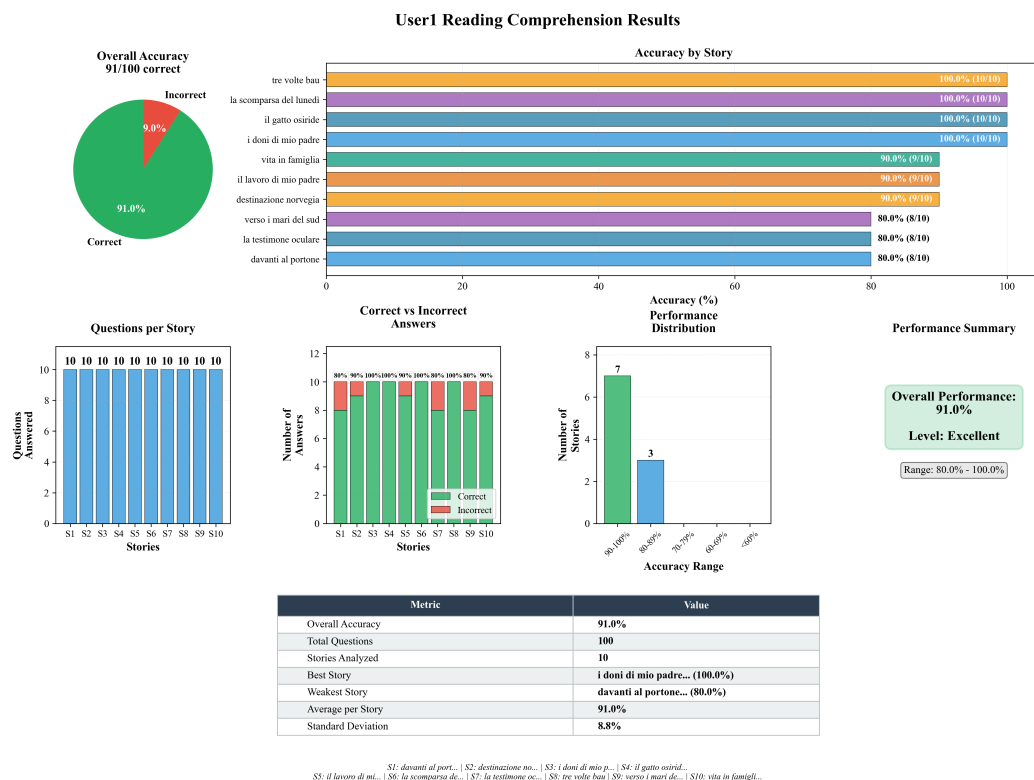


Figure 9: Example of an individualized reading-comprehension report generated for a participant. The upper-left panel summarizes overall comprehension accuracy, while the upper-right panel reports performance separately for each story. The lower panels summarize the number of questions administered per story, correct and incorrect responses, the distribution of accuracy values, and overall performance. Together, these views provide therapists with a compact overview of the participant's strengths and of texts that may require additional practice.



Figure 10: Example of the longitudinal performance report available to therapists. The upper-left panel shows comprehension accuracy across successive attempts together with the overall temporal trend. The upper-right panel compares mean performance during the early and late intervention periods. The lower-left panel summarizes repeated-question outcomes, distinguishing responses that remained correct, improved, remained incorrect, or worsened. The lower-right panel reports weekly average accuracy and highlights fluctuations in performance across the intervention.