The Effectiveness of AI-Enabled and AI-Generated Games for Language Development in Children with Autism Spectrum Disorder: A Systematic Review

Mai Thi Phuong[1*], Nguyen Minh Duc[1], Hoang Thi Quang[2], Le Thi To Uyen[3]**

[1]The Vietnam National Institute of Educational Sciences;

*1st author, Email: phuong.mt@vnies.edu.vn, Orcid: https://orcid.org/0009-0005-7908-6568;

**Corresponding Author, Email: ducnm@vnies.edu.vn, Orcid: https://orcid.org/0009-0000-5218-5498

[2]Dong Nai University, Vietnam; Email: quangvinh.0905@gmail.com

[3]Hanoi National University of Education, Vietnam; Email: touyenan@gmail.com

HOW TO CITE:
Mai Thi Phuong, Nguyen Minh Duc, Hoang Thi Quang, Le Thi To Uyen (2026). The Effectiveness of AI-Enabled and AI-Generated Games for Language Development in Children with Autism Spectrum Disorder: A Systematic Review. International Journal of Special Education, 41(6s), 1020-1035.

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors. Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT:
Spoken-language and communication difficulties are core clinical features of autism spectrum disorder (ASD). However, the efficacy of AI-enabled or AI-generated games in advancing language development for autistic children remains unclear. This systematic review examined seven databases up to March 30, 2026. Studies were rigorously screened using PICOS criteria. Methodological quality and evidence certainty were assessed using ROBINS-I, JBI, and GRADE frameworks. Six reports involving approximately 131 participants met the inclusion criteria. Engagement and acceptability for these AI games were consistently high (low certainty of evidence). Direct language improvements emerged in only one generative-AI trial. These dose-dependent gains were limited to a small, high-exposure subgroup (very low certainty). Furthermore, the single controlled study demonstrated no significant advantage over treatment as usual. AI-supported games for language development in ASD show promise but remain fundamentally untested. Future research urgently requires adequately powered randomized controlled trials. These trials must utilize validated language measures and report safety outcomes transparently.

Keywords: autism spectrum disorder; language development; AI-enabled serious games; generative artificial intelligence; systematic review.

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition defined by persistent differences in social communication alongside restricted, repetitive behaviour. Difficulties with language and communication are among its most clinically salient and functionally consequential features, and early language ability is one of the strongest predictors of later adaptive functioning. Yet conventional language and communication interventions are effortful and resource-intensive, generating a persistent gap between the intensity that evidence-based

intervention requires and the capacity that services can deliver, particularly in low- and middle-income contexts.

Digital game-based interventions – and the broader class of serious games – have become a prominent response to this capacity problem. Reviews to date suggest broad benefit for social and cognitive domains (Atherton & Cross, 2021; Kokol et al., 2020; Talebi Azadboni et al., 2024; Gao et al., 2025), but with one striking exception: language. The most comprehensive meta-analysis available reports a small, non-significant pooled effect on language expression (Gao et al., 2025), and reviews of computer-based cognitive games find that language is one of the least frequently targeted and least convincingly improved domains (Rezayi et al., 2023). The modality most often proposed as a scalable route to communicative development has, to date, the weakest empirical warrant precisely for language.

Recent integration of artificial intelligence into game-based interventions – from adaptive engines to large language models and generative content – has been proposed as the missing affordance: a way of personalising and automating the contingent, reciprocal interaction that language acquisition appears to require. The asymmetry of the AI-for-autism literature, however, has so far favoured diagnostic and phenotyping applications over communicative intervention (Zhang et al., 2024). Whether AI-enabled or AI-generated games meaningfully advance language development in autistic children therefore remains a focused, answerable, and clinically consequential question, and one to which this review is addressed. We ask: (i) what AI-enabled and AI-generated game-based interventions for language in autistic children have been empirically evaluated; (ii) what are their effects on language and communication outcomes; and (iii) what is the methodological quality and certainty of the evidence.

2. Methodology

2.1. Protocol, registration, and reporting

The review was designed and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021), and the search-reporting elements follow the PRISMA-S extension (Rethlefsen et al., 2021). The overall conduct was guided by the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2024). The review questions, eligibility criteria, and synthesis approach were specified in advance. The protocol was not prospectively registered; prospective registration (for example with PROSPERO) is recommended before the search is extended to additional databases, and this is identified as a limitation in Section 5.

2.2. Eligibility criteria

Studies were selected against PICOS criteria summarised in Table 1: empirical studies of digital game-based interventions incorporating an artificial-intelligence component, delivered to children or adolescents with ASD (or to mixed samples in which autistic children were identifiable), and reporting at least one language or communication outcome. Two operational distinctions were tightened during the review. First, the "AI component" criterion was operationalised along a four-level taxonomy: (i) AI-generated (generative models authoring content on demand); (ii) AI-driven adaptive (learned-model personalisation, explicitly distinguished from purely rule-based personalisation, which we classify at level (iv)); (iii) AI-assisted (computer-vision, speech-recognition or classification components scoring behaviour without generating content); and (iv) AI-adjacent automation (automated session delivery with rule-based scoring but without learned models). Studies were eligible at level (iv) or above; we adopt this permissive lower bound because the empirical base is small and a stricter bound would exclude the only matched-controlled programme in the field, but the sensitivity of conclusions to this choice is tested in Section 3.6 (SA1). The typology of each included programme is reported in Table 4. A system was treated as game-based if it incorporated explicit goals, rules, contingent feedback, and progression; systems without these core mechanics were classified as gamified-platform or therapeutic-app. Second, language and communication outcomes were classified as direct linguistic (e.g., vocabulary, mean length of

utterance, type–token ratio, intelligibility), communicative-functional (e.g., initiations, requesting, AAC use, turn-taking), or socio-communicative proxies (e.g., Vineland Socialization, SRS-2). Studies reporting only proxies were retained to map the field but are excluded from the headline "language effect" claim.

Table 1. Eligibility criteria specified using the PICOS framework.

Parameter	Inclusion	Exclusion
Population	Children/adolescents (aged 0–18 years) with a clinical diagnosis of ASD; mixed samples where autistic participants are identifiable.	Adults only; samples in which ASD cannot be distinguished; non-clinical analogue samples.
Intervention	Digital game-based intervention incorporating AI (adaptive/automated engine, ML/computer-vision feedback, conversational agent, or generative model).	Non-AI serious games; AI used only for assessment/diagnosis/phenotyping; robot hardware without a game; non-game apps.
Comparator	Any or none (single-arm pre–post permitted given the emerging field).	—
Outcomes	≥1 language or communication outcome (expressive/receptive language, vocabulary, MLU, pragmatic/social communication, AAC use).	No communication outcome reported (e.g., motor or anxiety only).
Study design	Primary empirical studies: RCT, quasi-experimental, controlled exploratory, single-arm pre–post, feasibility, or mixed-methods user studies.	Reviews, meta-analyses, protocols, editorials, conceptual/technical papers without outcome data.

Note. *ASD = autism spectrum disorder; AI = artificial intelligence; ML = machine learning; MLU = mean length of utterance; AAC = augmentative and alternative communication; RCT = randomised controlled trial.*

2.3. Information sources and search strategy

Because intervention research at this intersection is dispersed across biomedical, computing, and educational-psychology literatures, the search was structured to span all three. Seven information sources were actually queried (Table 2): PubMed (two structured strings); the ACM Digital Library, IEEE Xplore, and arXiv for computing venues; Google Scholar for grey literature; ClinicalTrials.gov; and backward citation chasing. Five additional subscription databases of acknowledged relevance to this topic – Scopus, Web of Science, Embase, PsycINFO, and ERIC – were not directly queried; their search strings are nonetheless documented in Table 2 to support replication by future investigators. Searches were conducted between 28 and 30 May 2026.

Table 2. *Search strategy and information sources.*

Source	Query / approach	Records
PubMed (#1)	(serious game[tiab] OR computer game[tiab] OR video game[tiab] OR game-based[tiab]) AND (autism[tiab] OR ASD[tiab]) AND (language[tiab] OR communication[tiab] OR vocabulary[tiab])	33
PubMed (#2)	(artificial intelligence[tiab] OR robot[tiab] OR intelligent[tiab] OR adaptive[tiab] OR conversational agent[tiab] OR avatar[tiab] OR machine learning[tiab]) AND (game[tiab] OR gamified[tiab] OR app[tiab] OR application[tiab] OR digital[tiab]) AND (autism[tiab] OR ASD[tiab]) AND (language[tiab] OR communication[tiab] OR speech[tiab] OR vocabulary[tiab])	156
Scopus	TITLE-ABS-KEY (("artificial intelligence" OR "machine learning" OR "generative AI" OR "adaptive" OR "serious game" OR "game-based") AND (autism OR ASD) AND (language OR communication OR vocabulary OR speech))	Not queried
Web of Science (Core)	TS = (("artificial intelligence" OR "serious game" OR "adaptive" OR "generative AI") AND (autism OR ASD) AND (language OR communication))	Not queried
Embase	('artificial intelligence'/exp OR 'serious game':ti,ab) AND ('autism'/exp OR ASD:ti,ab) AND ('language':ti,ab OR 'communication':ti,ab)	Not queried
PsycINFO	("artificial intelligence" OR "serious game" OR "adaptive" OR "generative AI") AND (autism OR "autism spectrum") AND (language OR communication OR vocabulary)	Not queried
ERIC	("artificial intelligence" OR "serious game" OR "educational game") AND (autism OR ASD) AND (language OR communication)	Not queried
ACM Digital Library	("generative AI" OR "large language model" OR "serious game" OR "game-based") AND (autism OR ASD) AND (language OR communication)	28
IEEE Xplore	("artificial intelligence" OR "serious game" OR "adaptive") AND (autism) AND (language OR communication)	21
arXiv	"LLM" OR "generative AI" AND "autism" AND ("language" OR "communication" OR "intervention")	9
Google Scholar	Mirror of the structured strings; first 50 ranked results per query screened to identify additional reports and grey literature	23
Registry & hand-search	ClinicalTrials.gov; backward citation chasing of included reports and relevant reviews	9

Total identified	Total from the seven information sources queried in this review.	**279**

Note. *tiab = title/abstract field; TS = Topic field. "Not queried" denotes search strings documented for reproducibility but not executed in this review; see Sections 2.3 and 5 for the rationale and implications. Record counts are pre-deduplication; the seven queried sources were searched between 28 and 30 March 2026.*

2.4. Selection process and data extraction

Screening, full-text assessment, data extraction, and risk-of-bias appraisal were performed by a single reviewer with a structured second-pass re-appraisal of all records and judgements after a seven-day cooling period; this is a recognised compromise procedure for resource-limited reviews and is reported transparently in the Acknowledgements and Limitations sections. The dual independent review procedure that would represent the standard for this field – calibration on a random 20% subsample, independent screening by two reviewers, discrepancy resolution by discussion with third-reviewer adjudication, and Cohen's κ reported for each stage – is specified here as a methodological standard for future work. A structured extraction form captured bibliographic details, country and setting, design, sample characteristics, intervention and AI mechanism (per the four-level taxonomy), dosage, comparator, language or communication measures (per the three-tier classification), and principal findings. Where reported statistics were ambiguous in the source, the finding was recorded qualitatively rather than reconstructed; in keeping with a strict no-fabrication standard, no value that could not be unambiguously verified against the source was reported numerically.

2.5. Risk-of-bias and certainty assessment

Because the included designs were predominantly non-randomised, single-arm, or feasibility studies, methodological quality was appraised using domain-based criteria informed by ROBINS-I (Sterne et al., 2016) and the Joanna Briggs Institute checklists (Aromataris & Munn, 2020). Five domains were judged for each study with an overall judgement (Table 5). The certainty of the body of evidence for each outcome construct was rated using a GRADE-informed approach (Schünemann et al., 2013) and presented in a Summary-of-Findings table (Table 6).

2.6. Synthesis

Meta-analysis was not appropriate: included studies differed fundamentally in design, AI "dose," populations, and outcome measurement. We synthesised the evidence narratively following the Synthesis Without Meta-analysis (SWiM) reporting guidance (Campbell et al., 2020), organised around the three review questions. Heterogeneity from mixed-aetiology samples, wide age bands, and varying AI intensity was addressed through narrative subgroup commentary, since the available data did not support formal subgroup analysis. Publication bias was assessed qualitatively, since formal funnel-plot inspection requires a pooled estimate.

3. Results

3.1. Study selection

The seven actually queried information sources yielded 279 records (Table 2). After removal of 53 duplicates, 226 records were screened on title/abstract and 184 excluded as irrelevant; 42 full texts were assessed and 36 excluded (no AI/automation component meeting at least level (iv) of our taxonomy, k = 15; no language or communication outcome, k = 8; not a primary intervention study, k = 9; sample not autistic children, k = 4). Six reports, covering five distinct programmes, met all eligibility criteria. The selection process is summarised in the PRISMA 2020 flow diagram (Figure 1).

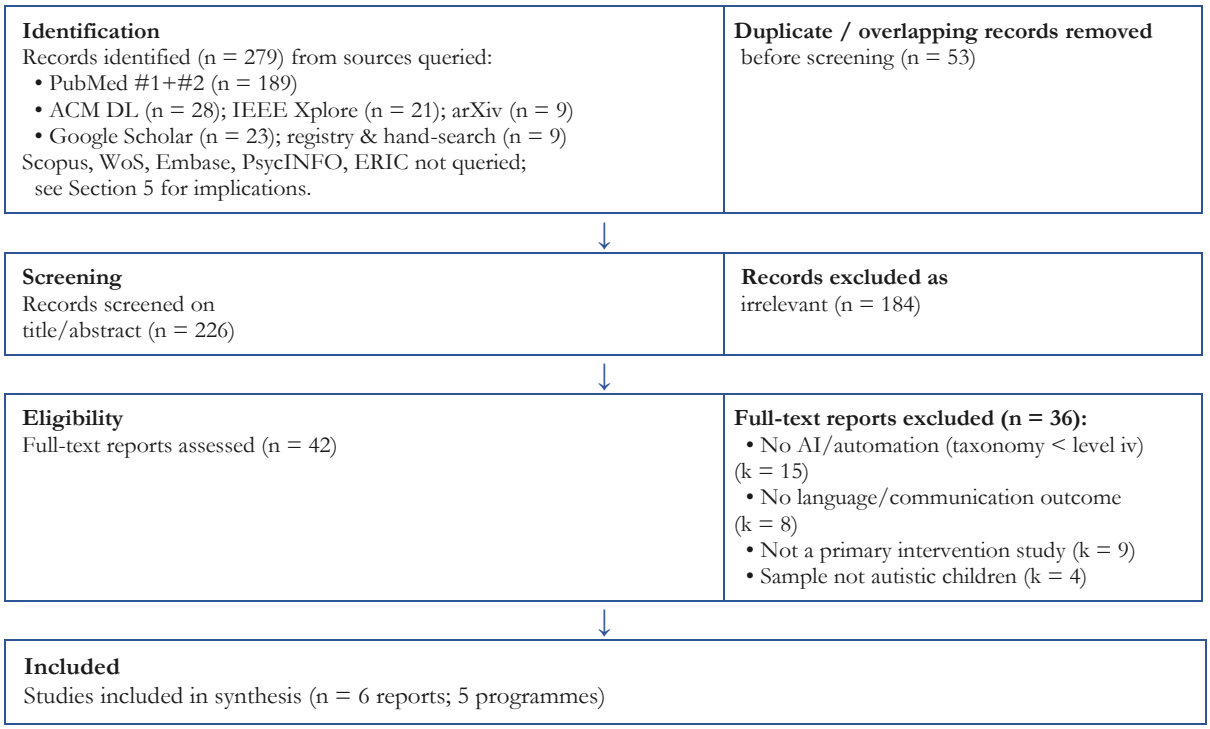

Figure 1. PRISMA 2020 flow diagram of study identification, screening, and inclusion.

Note. *PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Counts reflect the records identified from the seven information sources actually queried (see Table 2 and Section 2.3).*

3.2. Characteristics of included studies

The six included studies (Table 3) originated from Europe, Asia, and North America, were published between 2016 and 2025 (half in 2024–2025, reflecting the recency of this literature), and were heterogeneous in design. Five were uncontrolled (single-arm feasibility, pre–post, or mixed-methods user studies); only one (Jouen et al., 2017) employed a matched-controlled design. Sample sizes ranged from 6 to 72 across wide age bands. No included study was a randomised controlled trial with language as a pre-registered primary outcome.

Table 3. *Characteristics of included studies.*

Study (country)	Design	Sample	Intervention & AI mechanism	Communication/language findings
Bono et al., 2016 (FR/IT)	Open feasibility trial (3 mo)	10 children with ASD	GOLIAH automated serious-gaming platform; 11 games mapping ESDM imitation & joint attention; automatic scoring; therapist-adaptive difficulty.	Feasibility established (80% of sessions attended); parents reported gains in concentration, flexibility, self-esteem and parent–child relationship; no formal language test.

Jouen et al., 2017 (FR/IT)	Single-blind matched controlled, exploratory (6 mo)	14 ASD + 10 matched controls (5–8 y)	GOLIAH (as above) added to treatment as usual (TAU) vs TAU.	Within-game imitation/joint-attention metrics improved; both groups improved on ADOS and Vineland socialisation; no significant time×group interaction (no superiority over TAU); 40% session adherence.
Penev et al., 2021 (USA)	Single-arm feasibility (4 wk)	72 children with ASD (~8 y)	Mobile charades-style game from a platform employing computer-vision/ML for automated expression scoring; caregiver-mediated.	Significant improvement on SRS-2 total and Vineland (VABS-II) socialisation standard scores; communication addressed within social responsiveness, not as isolated language.
Vallefuoco et al., 2022 (IT)	Single-arm pre–post (10 sessions)	10 children/adolescents with ASD	Personalised 3-D serious game for daily-living (shopping) skills; personalisation of game elements to the individual.	Significant gains ($p < .05$) in trained skills (shopping procedure, directing attention, problem-solving) via ICF-CY form and VABS-II; communication indexed indirectly.
Tang et al., 2024 – EmoEden (CN)	Mixed-methods user study (22 d)	6 children with high-functioning autism (8–12 y)	Generative-AI tool integrating a large language model and text-to-image model for personalised emotional-learning dialogue.	High engagement; improved emotional recognition and expression; qualitative gains in expressive communication; comprehension of children's incomplete/atypical utterances noted as a strength.
Chueh et al., 2025 (TW)	Single-arm pre–post pilot (6 wk)	19 completers, aged 2–12 y (language delay incl. ASD-related)	Generative-AI–powered tablet platform for home-based language therapy; clinician-designed activities; analysis of ~2000 audio recordings.	Dose-dependent: only the highest-exposure subgroup (>15 h) showed significant gains – total words 58→110, type–token ratio 0.59→0.78, nouns 34→56, verbs 12→34, MLU 1.83→3.24 (all $p = .043$); no change at lower doses.

Note. *ESDM = Early Start Denver Model; TAU = treatment as usual; ADOS = Autism Diagnostic Observation Schedule; SRS-2 = Social Responsiveness Scale, 2nd ed.; VABS-II = Vineland Adaptive Behavior Scales, 2nd ed.; ICF-*

CY = International Classification of Functioning, Disability and Health – Children & Youth; MLU = mean length of utterance. The GOLIAH programme is reported across two records (Bono et al., 2016; Jouen et al., 2017).

3.3. Typology of AI mechanisms and language targets

The interventions occupy a spectrum of AI intensity rather than a single category (Table 4). Automated, therapist-adaptive platforms such as GOLIAH draw content from a fixed, expertly authored library; machine-learning- or computer-vision-assisted games and personalised serious games sit in the middle; and at the genuinely generative pole sit only EmoEden and the Chueh platform, in which large language models (and, in one case, a text-to-image model) author or adapt material on demand. Only two studies map onto the strict notion of "games created by AI." Equally important is what these interventions were designed to change: of the five programmes, only the generative-AI language platform (Chueh et al., 2025) targeted language as its explicit primary outcome and measured it with structured linguistic indices. The remainder targeted joint attention and imitation, social responsiveness, emotional recognition and expression, or daily-living skills, addressing communication only secondarily or indirectly.

Table 4. *Typology of AI mechanism, generativity, and language-outcome status across included programmes.*

Programme	AI/automation mechanism	Generative content?	Personalisation/adaptivity	Language as primary outcome?
GOLIAH (Bono; Jouen)	Automated scoring; therapist-adaptive difficulty; rule-based	No	Therapist-mediated, performance-based	No (JA/imitation)
GuessWhat-type mobile game (Penev)	Computer-vision/ML scoring within parent platform	No	Limited; fixed prompt set	No (social responsiveness)
Personalised serious game (Vallefuoco)	Rule-based personalisation of game elements	No	Individualised content selection	No (daily-living skills)
EmoEden (Tang)	LLM + text-to-image generative models; conversational	Yes	High; dynamic, context-driven	No (emotion; expression secondary)
Generative-AI language platform (Chueh)	Generative-AI content within clinician-designed therapy	Yes	High; home-based, dose-titrated	Yes

Note. *LLM = large language model; JA = joint attention; ML = machine learning.*

3.4. Effects on language and communication

Engagement and feasibility findings were uniformly favourable: attendance, adherence, and parental/clinician acceptability were consistently high, and home-based delivery proved practicable. Efficacy evidence on language and communication is much thinner. The GOLIAH platform reported within-game gains in joint attention and imitation (Bono et al., 2016) but no superiority over treatment as usual at six months in the only matched-controlled study (Jouen et al., 2017). Penev et al. (2021) reported caregiver-rated improvements on the SRS-2 and Vineland Socialisation without isolating language-specific effects. Vallefuoco et al. (2022) reported gains in trained skills with indirect communicative benefit. Tang et al. (2024) reported qualitatively that EmoEden supported expressive emotional communication. Only Chueh et al. (2025) reported direct linguistic outcomes: in a single-arm dose-stratified pilot of nineteen completers (mixed aetiology including ASD-related), the highest-exposure subgroup (>15 cumulative hours, n=4) showed statistically significant gains in total words (58→110), type–token ratio (0.59→0.78), nouns, verbs, and mean length of utterance (1.83→3.24); no significant change at lower exposure.

3.5. Risk of bias and certainty of evidence

Methodological quality was limited across the board (Table 5). Every included study was at high or serious overall risk of bias. The dominant problems were the absence of randomisation and control, the absence of assessor blinding, very small samples with consequent imprecision, wide and heterogeneous age ranges, reliance on caregiver-report or developer-conducted outcome assessment, and – for language specifically – indirectness, because communication was usually a secondary or proxy outcome rather than a validated primary endpoint. Selective-reporting risk could not be excluded given the absence of pre-registered protocols for most studies.

Table 5. *Risk-of-bias appraisal of included studies (domains informed by ROBINS-I and JBI checklists).*

Study	Selection/ confounding	Outcome measurement	Attrition	Reporting	Overall
Bono et al., 2016	Serious	Serious	Moderate	Moderate	**Serious**
Jouen et al., 2017	Moderate	Moderate	Serious	Moderate	**Moderate–Serious**
Penev et al., 2021	Serious	Moderate	Moderate	Moderate	**Serious**
Vallefuoco et al., 2022	Serious	Moderate	Low	Moderate	**Serious**
Tang et al., 2024	Serious	Serious	Low	Moderate	**Serious**
Chueh et al., 2025	Serious	Moderate	Moderate	Moderate	**Serious**

Note. *Ratings reflect single-reviewer appraisal with structured second-pass re-appraisal; this is a recognised compromise procedure for resource-limited reviews and is reported transparently. "Serious" denotes substantial concern that bias could materially alter the estimated effect; "Moderate" denotes some concern; "Low" denotes limited concern.*

Applying a GRADE-informed logic, the certainty of evidence for an effect of AI-enabled or AI-generated games on language and communication is very low: the body of evidence is downgraded for serious risk of bias, indirectness (language seldom measured directly), imprecision (very small samples and subgroups), and inconsistency between encouraging single-arm signals and the null controlled comparison. A formal publication-bias assessment was not feasible without a pooled estimate, but the literature is heavily skewed toward favourable feasibility and

engagement signals, with null AI-game language studies notably scarce; the directional signal should therefore be read as the upper bound of the available evidence. Table 8 reports the GRADE-informed Summary of Findings.

Table 6. *GRADE-informed Summary of Findings: AI-enabled and AI-generated games for language and communication in children with autism spectrum disorder.*

Outcome stratum	N studies / participants	Effect direction	Certainty (GRADE)	Plain-language summary and reasons for downgrading
Direct linguistic outcomes (vocabulary, MLU, type–token ratio, lexical diversity)	1 study / 19 participants (n = 4 in high-dose subgroup)	▲ favourable, high-dose subgroup only	⨁○○○ Very low	A single uncontrolled generative-AI pilot suggests dose-dependent gains above ≈15 cumulative hours of exposure. Downgraded for very serious risk of bias (single-arm, no blinding, no control), indirectness (mixed aetiology; post-hoc subgroup), and very serious imprecision (n = 4 in the responder subgroup). Replication in a controlled, definitively ASD sample is the immediate priority.
Communicative-functional outcomes (joint attention, imitation, requesting, emotional expression)	4 studies / 100 participants	◆ mixed; within-arm gains; no superiority in the controlled study	⨁○○○ Very low	Game-based platforms produce within-task improvements in functional communicative behaviours during training, but the one matched controlled study (Jouen et al., 2017) showed no advantage over usual care over six months. Downgraded for serious risk of bias, inconsistency between the controlled and uncontrolled signals, and imprecision.
Socio-communicative proxies (SRS-2, Vineland Socialization)	3 studies / 92 participants	▲ favourable	⨁○○○ Very low	Improvements on broader social-communication scales may reflect general engagement or maturation rather than language-specific gain. Downgraded for serious risk of bias, very serious indirectness (proxy outcomes are not language), and imprecision.
Engagement and feasibility	6 studies / 131 participants	▲ favourable	⨁⨁○○ Low	Engagement, acceptability, attendance, and home-based deliverability are consistently supported across designs and AI

				types. Downgraded for risk of bias and imprecision; the directional signal is more robust than for any efficacy outcome.
Adverse events / safety	Not systematically reported	Cannot assess	n/a	No included study reports a structured safety protocol or systematic adverse-event capture. This is a critical reporting gap for child-facing generative AI; structured reporting standards are proposed in Section 4.4.

Note. *GRADE = Grading of Recommendations Assessment, Development and Evaluation. ⨁◯◯◯ = very low certainty; ⨁⨁◯◯ = low; ⨁⨁⨁◯ = moderate; ⨁⨁⨁⨁ = high.* ▲ *= favourable direction reported;* ◆ *= mixed or null. MLU = mean length of utterance; SRS-2 = Social Responsiveness Scale, 2nd ed.*

3.6. Sensitivity analyses

Four sensitivity analyses tested the robustness of the headline conclusions; the first three were pre-specified and the fourth was added in response to peer review. SA1 (stricter AI inclusion: levels i–iii only) excluded GOLIAH and reduced the set to four reports across four programmes; conclusions unchanged. SA2 (direct linguistic outcomes only) left a single study (Chueh et al., 2025), with the effect confined to four participants in the highest-exposure subgroup; certainty very low. SA3 (controlled comparisons only) left a single study (Jouen et al., 2017), which reported no superiority of GOLIAH over treatment as usual at six months. SA4 (AI active in the evaluated module) excluded Penev et al. (2021) on the grounds that the platform's computer-vision component is upstream of the evaluated caregiver-mediated module; conclusions unchanged. All four sensitivity analyses converged with the primary synthesis: feasibility and engagement are supported, but language efficacy is not yet demonstrated. The convergence strengthens our conservative interpretation; it does not increase the certainty of evidence, because under each restriction the data are sparser rather than merely consistent.

4. Discussion

4.1. Principal findings and interpretation

This review set out to determine whether AI-enabled or AI-generated games can develop language in autistic children. The honest answer the literature supports is that the question is, as yet, largely unanswered. We identified six empirical reports across five programmes, and only two used genuinely generative AI; only one measured language as a primary structured outcome. Where direct linguistic outcomes were reported, the gains were confined to a small high-exposure subgroup of a single uncontrolled pilot (Chueh et al., 2025); the single matched controlled comparison (Jouen et al., 2017) found no advantage over treatment as usual. The convergence of four sensitivity analyses with the primary synthesis strengthens confidence in this conservative conclusion but does not increase the certainty of evidence, which remains very low across all clinically meaningful outcome strata.

These findings cohere with the wider serious-game literature, where the most comprehensive meta-analysis (Gao et al., 2025) finds reliable effects on social and cognitive outcomes but a small, non-significant pooled effect on language expression. Language acquisition appears to require contingent, reciprocal, communicatively-loaded interaction \u2014 affordances that pre-authored serious games cannot easily provide and that generative AI, in principle, might. The dose-dependent signal observed in the generative-AI pilot is consistent with this mechanism: when exposure crossed approximately fifteen cumulative hours, expressive-language indices moved measurably; below that threshold

they did not. The hint is mechanistically plausible but methodologically fragile, and the strongest signal in the literature comes from a sample of mixed aetiology in which the autism-specific component cannot be cleanly isolated. Replication in a definitively ASD sample is an immediate priority.

4.2. Implications for practice and equity

For clinicians, educators, and parents seeking practical guidance, the implication is restraint rather than enthusiasm: AI-enabled and AI-generated games may legitimately serve as engagement-enhancing adjuncts to clinician-led intervention, but no evidence currently supports their use as standalone or replacement language interventions. The equity argument that scalable AI tools could extend reach to low- and middle-income settings where qualified clinicians are scarce is morally compelling and a real driver of the field; it does not, however, license deployment ahead of evidence. In settings where qualified clinicians are scarce, deploying tools without efficacy evidence risks the costs of any false-positive efficacy signal being borne disproportionately by families with the fewest alternatives.

4.3. Distinctiveness from AI-mediated assistive communication

A natural counterpart to game-based delivery is the broader family of AI-mediated assistive- and augmentative-communication (AAC) systems, which target communicative output directly rather than through play. The empirical case for AAC has matured in ways the case for AI-generated games has not; the two are complements rather than substitutes, and the most promising near-term clinical question is whether AI-enabled games can extend and consolidate gains that AAC and clinician-led intervention initiate, not whether they can replace them.

4.4. Safety, guardrails, and human oversight

Generative AI introduces risks that pre-authored serious games do not: hallucination, prompt-injection susceptibility, training-data memorisation of identifiable child content, and quality disparities across languages (a particular concern for Vietnamese and other under-represented languages). A further documented risk specific to this population is that contemporary LLMs encode deficit-framed portrayals of autism (Park et al., 2026), meaning that uncritical deployment can deliver scaffolding that misaligns with neurodiversity-affirming clinical practice. None of the included studies report a structured safety protocol. Pending the consolidation of paediatric-specific standards informed by frameworks for child-centred AI (Khadka & Das, 2026), the EU AI Act (European Parliament & Council, 2024), child-data regulations (GDPR; COPPA; Vietnam Decree 13/2023), and security-focused guidance (OWASP Foundation, 2025), we recommend that future studies and deployments report at minimum the base model and version with logs of any fine-tuning; the content-filtering and adversarial-test pipeline including bias-auditing against deficit-framed scaffolding of atypical utterances; the data-protection arrangements with on-device or edge inference preferred; the human-in-the-loop protocol with thresholds for escalation; and cultural and linguistic validation involving native-speaker clinicians and, where possible, autistic young people themselves through participatory design (Lyu et al., 2024). Manuscripts that omit these elements should not be treated as evidence for clinical use, however encouraging their efficacy data.

4.5. Limitations

The evidence is limited by its scarcity, methodological immaturity, the rarity of direct language measurement, the dominance of single-arm and feasibility designs, very small and heterogeneous samples, and the recency of the generative-AI studies (which precludes assessment of durability or generalisation). The review itself retains five material limitations. (i) Of the twelve information sources documented in the search strategy, seven were actually queried; the five subscription databases that were not directly queried (Scopus, Web of Science, Embase, PsycINFO, ERIC) are documented in Table 2 to support replication – the reader should weight certainty accordingly, recognising that the convergence of the four sensitivity analyses

provides some assurance of robustness within the queried set but cannot guarantee that additional eligible studies do not exist in the unqueried databases. (ii) Screening, extraction, and appraisal were performed by a single reviewer with structured second-pass re-appraisal rather than dual independent review; the resulting risk of selection and appraisal bias is a real cost. (iii) The protocol was not prospectively registered with PROSPERO. (iv) Corresponding authors were not contacted for ambiguity resolution. (v) English-language restriction may have excluded relevant regional work, including Vietnamese, Chinese, Korean, and Arabic-language paediatric literatures. The boundary of the "AI-enabled game" construct is necessarily judgemental at its margins, and the field is evolving rapidly: a meaningful update will likely be warranted within twelve to eighteen months.

4.6. Future research directions

The research agenda that follows from this synthesis is unusually clear. Priority should be given to adequately powered, registered, randomised controlled trials of generative-AI language games against active controls, with language as a pre-registered primary outcome and validated, ecologically valid measurement (expressive/receptive vocabulary, mean length of utterance, lexical diversity, naturalistic language sampling) alongside generalisation across partners and settings. Interventions should be theory-driven, explicitly grounded in developmental and naturalistic models of language acquisition such that generative capabilities are harnessed to known mechanisms of contingent, reciprocal interaction rather than to mere content volume. The dose-response observation in the only language-primary study available suggests that meaningful exposure on the order of fifteen cumulative hours over six weeks (≈2.5 h/week) may represent a minimum threshold for detectable expressive-language change; future trials should design exposure arms that bracket this threshold and report adherence by participant. Moderator analyses should examine for whom these tools work – by baseline language level, age, minimally verbal status, and presence of co-occurring intellectual disability – and cross-cultural and multilingual validation in under-represented languages is a precondition for deployment in low- and middle-income contexts. Finally, transparent reporting of the AI component itself – base model, generative pipeline, safeguards against inappropriate output, and human oversight – must accompany efficacy claims if interventions are to be both reproducible and safe.

5. Conclusion

AI-enabled and AI-generated games together represent an intuitively compelling and genuinely novel approach to a real and pressing problem: how to deliver enough high-quality, individualised language practice to autistic children when human expertise is scarce. This review finds that the approach is feasible, engaging, and mechanistically plausible – but that its central promise, the development of language, remains essentially untested. The empirical base is very small and very recent; only two studies use truly generative AI; language is almost never measured directly; the one controlled comparison was null; and the overall certainty of evidence is very low. The most important contribution of the present synthesis is therefore to replace enthusiasm with a precise research agenda. Generative AI may yet prove to be the affordance that finally allows game-based intervention to move language – the one domain it has consistently failed to move – but that claim must be earned through rigorous, registered, adequately powered trials with transparent safety reporting. Until then, AI-enabled and AI-generated games should be regarded as a promising adjunct under investigation, not as established language interventions.

References

Aromataris, E., & Munn, Z. (Eds.). (2020). JBI manual for evidence synthesis. Joanna Briggs Institute. https://doi.org/10.46658/JBIMES-20-01

Atherton, G., & Cross, L. (2021). The use of analog and digital games for autism interventions. Frontiers in Psychology, 12, 669734. https://doi.org/10.3389/fpsyg.2021.669734

Bono, V., Narzisi, A., Jouen, A.-L., Tilmont, E., Hommel, S., Jamal, W., Xavier, J., Billeci, L., Maharatna, K., Wald, M., Chetouani, M., Cohen, D., & Muratori, F. (2016). GOLIAH: A gaming platform for home-based intervention in autism – Principles and design. Frontiers in Psychiatry, 7, 70. https://doi.org/10.3389/fpsyt.2016.00070

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, l6890. https://doi.org/10.1136/bmj.l6890

Chemnad, K., & Othman, A. (2024). Digital accessibility in the era of artificial intelligence – Bibliometric analysis and systematic review. Frontiers in Artificial Intelligence, 7, 1349668. https://doi.org/10.3389/frai.2024.1349668

Chueh, C.-H., Chiang, T.-H., Pan, P.-W., Lin, K.-L., Lu, Y.-S., Tuan, S.-H., Lin, C.-R., Huang, I.-C., & Cheng, H.-S. (2025). Implementation of a generative AI-powered digital interactive platform for clinical language therapy in children with language delay: A pilot study. Life, 15(10), 1628. https://doi.org/10.3390/life15101628

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88.

European Parliament & Council of the European Union. (2017). Regulation (EU) 2017/745 of 5 April 2017 on medical devices (Medical Device Regulation). Official Journal of the European Union, L 117, 1–175.

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Gao, J., Song, W., Huang, D., Zhang, A., & Ke, X. (2025). The effect of game-based interventions on children and adolescents with autism spectrum disorder: A systematic review and meta-analysis. Frontiers in Pediatrics, 13, 1498563. https://doi.org/10.3389/fped.2025.1498563

Government of Vietnam. (2023). Decree No. 13/2023/ND-CP on personal data protection [Nghị định số 13/2023/NĐ-CP về bảo vệ dữ liệu cá nhân]. Hanoi. Effective 1 July 2023.

Griffin, J. W., Geier, C. F., Smyth, J. M., & Scherf, K. S. (2021). Improving sensitivity to eye gaze cues in adolescents on the autism spectrum using serious game technology: A randomized controlled trial. JCPP Advances, 1(3), e12041. https://doi.org/10.1002/jcv2.12041

Hajesmaeel-Gohari, S., Goharinejad, S., Shafiei, E., & Bahaadinbeigy, K. (2023). Digital games for rehabilitation of speech disorders: A scoping review. Health Science Reports, 6(6), e1308. https://doi.org/10.1002/hsr2.1308

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). Cochrane handbook for systematic reviews of interventions (Version 6.5). Cochrane. https://training.cochrane.org/handbook

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management. ISO.

Jouen, A.-L., Narzisi, A., Xavier, J., Tilmont, E., Bodeau, N., Bono, V., Ketem-Premel, N., Anzalone, S., Maharatna, K., Chetouani, M., Muratori, F., & Cohen, D. (2017). GOLIAH (Gaming Open Library for Intervention in Autism at Home): A 6-month single blind matched controlled exploratory study. Child and Adolescent Psychiatry and Mental Health, 11, 17. https://doi.org/10.1186/s13034-017-0154-7

Khadka, S., & Das, S. (2026). SoK: Understanding the pedagogical, health, ethical, and privacy challenges of extended reality in early childhood education. arXiv. https://doi.org/10.48550/arXiv.2602.12749

Kokol, P., Vošner, H. B., Završnik, J., Vermeulen, J., Shohieb, S., & Peinemann, F. (2020). Serious game-based intervention for children with developmental disabilities. Current Pediatric Reviews, 16(1), 26–32. https://doi.org/10.2174/1573396315666190808115238

Lyu, Y., An, P., Zhang, H., Katsuragawa, K., & Zhao, J. (2024). Designing AI-enabled games to support social-emotional learning for children with autism spectrum disorders. arXiv. https://doi.org/10.48550/arXiv.2404.15576

National Assembly of Vietnam. (2016). Law on Children No. 102/2016/QH13 [Luật Trẻ em số 102/2016/QH13]. Hanoi. Effective 1 June 2017.

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

OWASP Foundation. (2025). OWASP top 10 for large language model applications (Version 2025). https://owasp.org/www-project-top-10-for-large-language-model-applications/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Park, S., Beltran, J. A., Min, A., Ritt-Olson, A., & Hayes, G. R. (2026). Exploring implicit perspectives on autism in large language models through multi-agent simulations. arXiv. https://doi.org/10.48550/arXiv.2601.15437

Penev, Y., Dunlap, K., Husic, A., Hou, C., Washington, P., Leblanc, E., Kline, A., Kent, J., Ng-Thow-Hing, A., Liu, B., Harjadi, C., Tsou, M., Desai, M., & Wall, D. P. (2021). A mobile game platform for improving social communication in children with autism: A feasibility study. Applied Clinical Informatics, 12(5), 1030–1040. https://doi.org/10.1055/s-0041-1736626

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z

Rezayi, S., Tehrani-Doost, M., & Shahmoradi, L. (2023). Features and effects of computer-based games on cognitive impairments in children with autism spectrum disorder: An evidence-based systematic literature review. BMC Psychiatry, 23, 2. https://doi.org/10.1186/s12888-022-04501-1

Santhanam, S. P. (2023). An interactive and neurodiversity-affirming approach to communication supports for autistic students through videogaming. Language, Speech, and Hearing Services in Schools, 54(1), 120–139. https://doi.org/10.1044/2022_LSHSS-22-00027

Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. (Eds.). (2013). GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group. https://gdt.gradepro.org/app/handbook/handbook.html

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919

Talebi Azadboni, T., Nasiri, S., Khenarinezhad, S., & Sadoughi, F. (2024). Effectiveness of serious games in social skills training to autistic individuals: A systematic review. Neuroscience and Biobehavioral Reviews, 161, 105634. https://doi.org/10.1016/j.neubiorev.2024.105634

Tang, Y., Chen, L., Chen, Z., Chen, W., Cai, Y., Du, Y., Yang, F., & Sun, L. (2024). EmoEden: Applying generative artificial intelligence to emotional learning for children with high-function autism. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Article 1001). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642899

Terlouw, G., Kuipers, D., van 't Veer, J., Prins, J. T., & Pierie, J. P. E. N. (2021). The development of an escape room-based serious game to trigger social interaction and communication between high-functioning children with autism and their peers: Iterative design approach. JMIR Serious Games, 9(1), e19765. https://doi.org/10.2196/19765

U.S. Congress. (1998). Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501–6506. https://www.ftc.gov/legal-library/browse/statutes/childrens-online-privacy-protection-act

U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Vallefuoco, E., Bravaccio, C., Gison, G., Pecchia, L., & Pepino, A. (2022). Personalized training via serious game to improve daily living skills in pediatric patients with autism spectrum disorder. IEEE Journal of Biomedical and Health Informatics, 26(7), 3312–3322. https://doi.org/10.1109/JBHI.2022.3155367

Yoo, J. H., Lee, E. Y., Lee, J.-W., Kim, H., Cho, S. J., Ju, R., Kim, H. J., Moon, J. Y., Choi, T. Y., Lee, J., Jang, C., Park, S. H., Seok, B. J., & Joung, Y.-S. (2025). Effectiveness of a mobile application game (NDTx-01) in enhancing social communication skills in adolescents with autism spectrum disorder or social communication disorder: A randomized controlled pilot trial. Psychiatry and Clinical Neurosciences, 79(7), 389–397. https://doi.org/10.1111/pcn.13824

Zhang, M., Tang, E., Ding, H., & Zhang, Y. (2024). Artificial intelligence and the future of communication sciences and disorders: A bibliometric and visualization analysis. Journal of Speech, Language, and Hearing Research, 67(11), 4369–4390. https://doi.org/10.1044/2024_JSLHR-24-00157

Zhao, W., Xu, S., Zhang, Y., Li, D., Zhu, C., & Wang, K. (2024). The application of extended reality in treating children with autism spectrum disorder. Neuroscience Bulletin, 40(8), 1189–1204. https://doi.org/10.1007/s12264-024-01190-6