

Methods, Metrics, and Taxonomies for Evaluating Accessibility and Usability in Digital Learning Systems for Deaf and Hard-of-Hearing Users: Implications for Inclusive Engineering and Technical Education

Diana Effendi*, Eddy Soeryanto Soegoto, Bobi Kurniawan, Adam Mukharil Bachtiar

Universitas Komputer Indonesia, Bandung, Indonesia
diana.75725001@mahasiswa.unikom.ac.id

Abstract—Digital learning systems are increasingly used to support inclusive technical education for deaf and hard-of-hearing (DHH) students, yet evidence regarding their accessibility and usability remains fragmented. This study employed a PRISMA-based Systematic Literature Review (SLR), supported by bibliometric analysis, quality assessment, and thematic synthesis. The findings show that evaluation practices remain dominated by technical performance, while accessibility assessment, direct user involvement, and learning-outcome evaluation are less consistently reported. Based on these patterns, an Evaluation Evidence Taxonomy and evidence-gap map were developed and subsequently synthesised into a Design Knowledge Taxonomy. These outputs integrate technical performance, accessibility, usability, user evidence, context of use, and learning outcomes within a structured evaluation perspective. The resulting taxonomies provide an evidence-based foundation for evaluating digital learning systems intended to support inclusive technical education and serve as intermediate research artefacts for subsequent QA Model development.

Keywords—accessibility, DHH, digital learning systems, engineering education, evidence taxonomy, quality assurance, technical education, usability.

I. INTRODUCTION

Advances in digital learning technologies have expanded opportunities for inclusive learning in engineering and technical education. These developments are particularly important for deaf and hard-of-hearing (DHH) technical students, who may face barriers when accessing information and interacting with digital learning systems. Accessible and inclusively designed digital learning materials can help accommodate the learning needs of students with hearing impairment when accessibility is considered as part of the instructional design (Musayaro et al., 2023). Reliance on audio-based information, limited captioning, and inadequate sign language support can reduce the inclusivity of technology-mediated learning experiences (Ahn et al., 2025; Escudeiro et al., 2022; Lynch et al., 2026; Pérez-Arteaga et al., 2025; Reyes et al., 2023). In technical education, students increasingly interact with digital

platforms, simulations, multimedia, virtual environments, and other technology-based instructional resources. Therefore, limitations in accessibility and usability may directly affect participation, comprehension, and engagement in technical learning activities. Inclusive digital learning systems are consequently important for providing equitable learning opportunities for DHH technical students.

In engineering and technical education, the quality of a digital learning system cannot be assessed solely on the basis of functional completeness or technological sophistication. Technical students must also be able to perceive information, operate the system, understand its content, and interact effectively with digital learning resources. In engineering education, students' understanding also depends on how technical information and terminology are structured and explained, particularly when complex concepts must be interpreted systematically (Nandiyanto & Yulistiani, 2024). Accessibility, usability, and user experience (UX) are therefore central dimensions of educational technology quality. ISO 9241-11 emphasises effectiveness, efficiency, and satisfaction within the context of use, whereas the Web Content Accessibility Guidelines (WCAG) establish perceivability, operability, understandability, and robustness as fundamental principles of digital accessibility. These principles are particularly relevant in technical learning environments that rely on visual information, multimedia, interactive systems, simulations, and other technology-intensive resources.

Multimedia, captions, augmented reality, virtual reality, and artificial intelligence have increasingly been used to reduce communication barriers for DHH learners. These technologies are also relevant to engineering and technical education because technical learning increasingly involves digitally mediated demonstrations, simulations, visualisation, interactive environments, and technology-assisted communication. However, technological innovation has not always been accompanied by equally rigorous educational evaluation. Some studies emphasise accuracy, precision, recall, or latency, whereas others employ usability

instruments, user evaluations, or specialised caption-testing methods. Formal accessibility audits and measurements of learning outcomes remain less consistently reported. This kind of imbalance is very important in technical education, since high algorithmic or computational performance may not guarantee accessibility, usability, and educational effectiveness of the system for technical learners.

This state of affairs suggests that systematic mapping needs to consider not only which technologies have been created but also the nature of evaluation evidence generated, stakeholders engaged, standards followed, and contexts of use (Ahn et al., 2025; Ariya et al., 2025; Berrezueta-Guzman & Wagner, 2025; Cabanillas-Carbonell et al., 2022; Loaiza-Ramírez and González-Beltrán, 2026; Pérez-Arteaga et al., 2025). In engineering and technical education, this distinction is important because the educational value of digital technology depends on the relationship between technical performance and actual learning use. The increasing research attention given to engineering education also demonstrates the importance of systematically connecting educational practices, technological development, and research evidence (Susilawati, 2024). Evidence regarding accessibility, usability, user participation, context of use, and learning outcomes is therefore required to determine whether digital systems can meaningfully support inclusive technical learning.

Given these circumstances, a systematic literature review (SLR) is required to map the structure of available evaluation evidence rather than merely summarising publication trends. PRISMA is used to ensure transparency in the study-selection process, whereas data extraction, quality assessment, thematic synthesis, taxonomy development, and evidence-gap mapping are used to compare evaluation methods, metrics, instruments, standards, DHH user involvement, contexts of use, learning outcomes, and the comprehensiveness of available evidence. The resulting structured evidence is subsequently synthesised into a Design Knowledge Taxonomy whose categories remain traceable to their underlying evidence sources. For engineering and technical education, this evidence-oriented approach provides a basis for evaluating whether digital learning technologies support technical students beyond technical performance alone.

This study aims to map the methods, metrics, and forms of evidence used to evaluate accessibility and usability in digital learning systems for deaf and hard-of-hearing (DHH) users; develop an Evaluation Evidence Taxonomy and an evidence-gap map; and derive a Design Knowledge Taxonomy through design-oriented synthesis. The study contributes two interrelated outputs: evidence mapping derived from the reviewed studies and design-oriented knowledge derived from the synthesis of that evidence. In terms of engineering and technical education, such outcomes show how technical performance, accessibility, usability, user evidence, context of use, and learning outcomes could be considered in the evaluation of digital learning systems for the purpose of supporting technical learners who are deaf and hard of hearing. Such outcomes are considered as

intermediate research outputs in the process of developing the QA Model. The research questions of the current study are:

- (i) RQ1: What trends, evaluation dimensions, methods, metrics, instruments, and standards are reported in studies evaluating accessibility and usability in digital learning systems for DHH users?
- (ii) RQ2: How could the provided evaluation evidence be structured into an Evaluation Evidence Taxonomy to allow comparison based on technical performance, accessibility, usability, user evidence, context of use, learning, and evidence integration?
- (iii) RQ3: What evidence gaps exist, and how can these gaps be addressed by developing a Design Knowledge Taxonomy that can further be used to develop a QA model for digital learning systems?

These results will be used to create a classification framework, which will allow for comparing evaluation approaches and linking scattered evidence with initial design knowledge, which is still traceable to its source. In engineering and technical education, such a framework will enable an assessment of digital learning technologies that is more well-rounded, where technology performance is separated from accessibility, usability, evidence of use by the user, context of use, and learning efficiency. The further development and validation of the QA Model are beyond the scope of the current research.

II. LITERATURE REVIEW

2.1. Quality of Digital Learning Systems in Engineering and Technical Education

The quality of digital learning systems utilized for the purpose of engineering and technical education depends on more than just the technical aspects of the system but on the ability of technical students to use the systems effectively. It is even more important when the target group consists of DHH learners due to the increased use of technological means and methods in technical learning. Therefore, the quality of such systems should include not only technical aspects but accessibility as well.

Three approaches can be applied to the evaluation of the quality of digital learning systems: accessibility, usability, and user experience (UX). Accessibility implies the ability of the system to meet the needs of various users and can be evaluated using criteria such as the Web Content Accessibility Guidelines (WCAG). Usability emphasises effectiveness, efficiency, and satisfaction within a specific context of use, whereas UX concerns users' perceptions of ease, comfort, engagement, and overall interaction with a system. These dimensions are particularly relevant to engineering and technical education because technological sophistication alone does not ensure meaningful technical learning. A digital system may function correctly while remaining difficult for DHH technical students to perceive, operate, understand, or use effectively. In this study, accessibility, usability, and UX are therefore treated as

analytical perspectives for organising evaluation evidence rather than as components of a new model.

2.2. Technology-Enhanced Learning for Deaf and Hard-of-Hearing Users

Digital technologies for DHH users have developed along two major directions that are relevant to inclusive technical education. Multimedia-supported learning has also been used in other instructional contexts to combine digital presentation with active learning activities (Abbood, 2024). The first focuses on improving access to learning content and learning experiences through technologies such as inclusive massive open online courses, adaptive captioning, augmented reality, and virtual reality (Ariya et al., 2025; Berrezueta-Guzman & Wagner, 2025; Cabanillas-Carbonell et al., 2022; Heck & Ferraro, 2025; Lynch et al., 2026; Reyes et al., 2023; Sivunen, 2025). These technologies provide alternative ways of presenting information and supporting interaction, which can help reduce communication barriers in digitally mediated technical learning environments. Digital media have also been applied to support the remote teaching of technical subject matter, illustrating their potential to extend access to instructional content beyond conventional classroom interaction (Ariyanti & Nandiyanto, 2022).

The second direction relates to the assistive technologies utilizing deep learning, computer vision, convolutional neural networks, transformers, and other computing methods for recognition and translation of sign language (Al-Saidi et al., 2024; Babalian & Slutu, 2025; Coster et al., 2024; González-Rodríguez et al., 2024; Hashi et al., 2025; Kim et al., 2025; Li et al., 2024; Nedjar & Mohammed, 2024; Panopoulos et al., 2025; Perea-Trigo et al., 2025; Rastgoo et al., 2021; Shahin & Ismail, 2024; Shin et al., 2023; Tan et al., 2024; Violet & Sri, 2025). Such innovations prove the increased use of computing technologies in learning and communication systems. The advent of new technologies has given rise to some opportunities and challenges for the delivery of technical education, stressing the necessity to integrate technological and educational requirements (Ogundele et al., 2025).

In terms of engineering and technical education, however, technological progress must be separated from educational efficiency. Digital transformation in education might extend learning possibilities, but the educational significance of this transformation is contingent on how technology is utilized for learning purposes (Al Husaeni & Rahmat, 2023). In technical learning, learners have to deal with complicated digital technologies, but a high level of algorithm efficiency, recognition efficiency, or speed of process cannot guarantee that technology is usable by learners and efficient for learning processes. Hence, evaluation should include technical characteristics as well as learners' ability to engage with technology and use it for learning processes.

2.3. Evaluation Approaches and Evidence Gaps

Even though there is significant technological advancement, it has been noted that the current literature

suffers from an imbalance in the kinds of evidence gathered to assess digital learning systems. For instance, many technology-based papers focus on measures of model accuracy, error rates, latency and other computer performance measures rather than user participation, usability assessments, accessibility testing, and learning performance (Ahn et al., 2025; Pérez-Arteaga et al., 2025).

Such imbalance is particularly significant in engineering and technical education, where a high level of technological performance does not necessarily imply that DHH technical students have access to information, are able to participate in learning activities and utilize the system for its educational purpose. Evaluation in technical education should therefore distinguish evidence of technological performance from evidence of accessibility, usability, user experience, and learning effectiveness.

The literature also varies considerably in reporting DHH participant characteristics and contexts of use, which limits comparison across studies. A comprehensive evaluation should therefore consider the methods employed, types of evidence generated, level of standardisation, extent of user involvement, context of use, and integration of different forms of evidence (Al-Saidi et al., 2024; Loaiza-Ramírez and González-Beltrán, 2026; González-Rodríguez et al., 2024; Hashi et al., 2025; Millett, 2021; Panopoulos et al., 2025; Perea-Trigo et al., 2025; Rastgoo et al., 2021; Shahin & Ismail, 2024; Shin et al., 2023; Sivunen, 2025; Tan et al., 2024; Violet & Sri, 2025).

For engineering education, such multidimensional evaluation is necessary because digital technologies should support both technological functionality and meaningful technical learning. An evidence structure integrating technical performance, accessibility, usability, user-based evidence, context of use, and learning outcomes can therefore provide a stronger basis for evaluating digital learning systems intended to support DHH technical students.

2.4. Human-AI Co-Creation as a Direction for Future QA Model Development

Generative artificial intelligence based on Large Language Models (LLMs) may support subsequent stages of quality assurance model development after research evidence has been systematically collected, evaluated, and organised. LLMs should not replace literature screening, data extraction, quality assessment, or evidence synthesis because their performance in systematic review activities remains variable and human reviewers continue to play an essential role in ensuring methodological reliability (Clark et al., 2025; Khraisha et al., 2024).

In subsequent research, the structured evidence generated in this study may provide input for Human-AI Co-Creation. LLMs may assist researchers in mapping relationships among evidence elements, exploring alternative structures, comparing conceptual representations, and examining conceptual consistency. However, a human-in-the-loop approach remains necessary so that interpretation, selection, correction, acceptance, and rejection remain under researcher control (Shah, 2025).

This approach has potential relevance to engineering and technical education because the development of inclusive digital learning systems requires technical, educational, accessibility, and user-centred considerations to be integrated. Human–AI collaboration may assist in organising complex evidence, but decisions concerning accessibility, usability, learning relevance, and the needs of technical students must remain grounded in empirical evidence and human judgement. Accordingly, LLMs are positioned outside the proposed QA Model and are considered only as potential design-support tools for subsequent model development. Human–AI Co-Creation is not implemented or evaluated in the present study and is therefore treated solely as a direction for future research.

III. METHOD

3.1. Systematic Literature Review and Study Selection

The study employed a Systematic Literature Review (SLR) in accordance with the PRISMA 2020 guidelines to ensure transparency and traceability throughout the study selection and literature analysis processes. The literature search was conducted using the Scopus database and covered publications from 2021 to 2025. The search terms encompassed three main domains: (1) evaluation, including ‘evaluation’, ‘assessment’, ‘analysis’, ‘quality’, ‘usability’, ‘accessibility’, ‘UX’, and ‘HCI’; (2) target users, including ‘deaf’, ‘DHH’, ‘hearing impairment’, ‘hard of hearing’, and ‘sign language’; and (3) learning contexts, including ‘learning’, ‘education’, ‘e-learning’, ‘digital learning system’, and ‘multimedia’. The study selection process is presented in Figure 1.

In the identification step, Scopus database searching was performed using an already determined set of keywords. The search was limited to the period between 2021 and 2025, the disciplines of Computer Science, Engineering, Social Sciences, and types of documents including research articles and conference papers and the language of the article as English. This resulted in 193 documents. Metadata used for this purpose included titles, authors, publication year, source/publisher, keywords, and Digital Object Identifier (DOI).

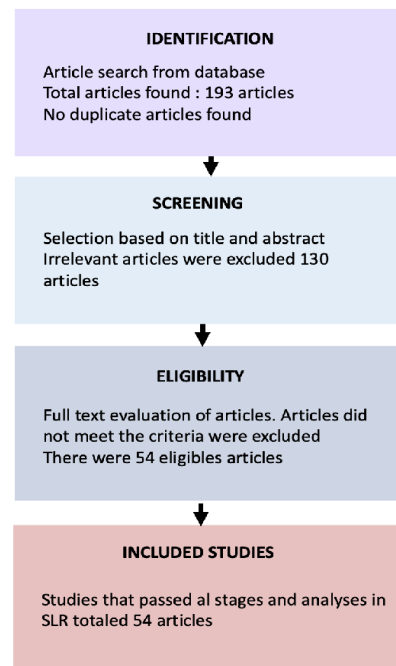


Fig. 1. PRISMA stages.

During the screening stage, all 193 records were assessed based on their titles and abstracts. No duplicate records were identified. A total of 139 documents were excluded because they did not align with the research objectives, including medical studies, technical studies without a clear learning context, and studies that did not focus on deaf and hard-of-hearing (DHH) users. Consequently, 54 studies proceeded to the full-text assessment stage.

During the eligibility stage, the full texts of the 54 studies were assessed to determine their alignment with the study objectives, methods, and context. Studies were included if they addressed digital learning systems, learning media, or learning software for deaf and hard-of-hearing (DHH) users and reported quality evaluations in the form of technical evidence, user-based evidence, accessibility evidence, learning outcomes, or a combination thereof. Information that was not available in the article was coded as NR (Not Reported).

A study was considered eligible if it met all of the following criteria:

- (i) The study addressed a digital learning system, learning medium, or learning software.
- (ii) The system or learning medium was intended for deaf and hard-of-hearing (DHH) users.
- (iii) The study reported an evaluation of the system’s quality, including technical performance, usability/UX, accessibility, learning outcomes, or a combination thereof.

Studies were excluded if they met any of the following criteria:

- (i) The research focused purely on technical aspects without a clear link to the learning context or the system being evaluated.
- (ii) The study did not involve an identifiable learning context.

- (iii) The study did not explicitly involve or target deaf and hard-of-hearing (DHH) users.

All 54 studies that qualified for full-text review were found to be eligible for inclusion in the study. Therefore, a total of 54 studies out of the total 193 documents screened were selected as part of the final analysis after excluding 139 studies.

3.2. Data Extraction, Quality Assessment and Evaluation Mapping

Information was gleaned from each of the selected articles on the kind of system used, user features, evaluation processes, evaluation tools/measures used, standards/frameworks cited, context of use, inclusion of deaf/hard-of-hearing users, learning outcomes, evidence integration and important findings. Data extraction was conducted for all 54 studies and served as the basis for the thematic synthesis and for addressing RQ1–RQ3.

The methodological quality of each study was assessed using the QA1–QA8 checklist with a binary scoring system (0 = absent or unclear; 1 = present or clear). The quality assessment was conducted to evaluate the completeness of reporting and methodological rigour and to inform the interpretation of the findings, rather than to serve as an exclusion criterion. The scores for the eight items were then summed to produce a total quality score ranging from 0 to 8.

QA1: The objectives of the usability/accessibility/UX evaluation are clearly explained (0 = absent/unclear; 1 = present/clear).

QA2: The context of use/learning is described (0 = absent/unclear; 1 = present/clear).

QA3: The target audience of people with hearing and speech impairments is clearly explained, including their criteria or characteristics (0 = absent/unclear; 1 = present/clear).

QA4: The evaluation methods are explained, including the procedures or steps involved (0 = none/unclear; 1 = present/clear).

QA5: The evaluation instruments or metrics are described, for example, SUS, USE, WCAG, or other technical metrics (0 = none/unclear; 1 = present/clear).

QA6: The sample/participants are reported, including their number and selection criteria (0 = none/unclear; 1 = present/clear).

QA7: Evaluation results are clearly reported in the form of figures, themes, or observational findings (0 = none/unclear; 1 = present/clear).

QA8: The study's limitations are stated (0 = none/unclear; 1 = present/clear).

A quality assessment was conducted for all 54 studies for which the full text was accessible in PDF or HTML format. The mean QA score was 5.93, with a median score of 6 and a range of 3–8. Of the 54 studies, 34 (63.0%) were classified as high quality (scores of 6–8), while 20 (37.0%) were classified as moderate quality (scores of 3–5). Regarding access format, 44 studies were available in PDF format and 10 through HTML/web-based formats. All studies were assessed based on the information reported in their full texts.

3.3. Development of the Evaluation Evidence Taxonomy and Evidence-Gap Map

The analysis was conducted through four sequential layers. First, bibliometric analysis was used to map the research landscape. Second, thematic synthesis and evaluation mapping were conducted by the researchers to identify patterns in methods, instruments, standards, user involvement, contexts of use, and types of evidence. Third, the mapping results were organised into the Evaluation Evidence Taxonomy and the evidence-gap map. Fourth, these two outputs were used in a design-oriented synthesis to derive categories of knowledge relevant to the QA Model formulation stage. Each category was required to be traceable to a dimension of the taxonomy, the evidence-gap map, or the standards and contexts analysed. This stage did not utilise LLMs and did not generate new empirical evidence.

3.4. Design-Oriented Synthesis

Design-oriented synthesis was conducted once the Evaluation Evidence Taxonomy and the evidence-gap map had been established. The researchers compared patterns and relationships across evaluation targets, methods and instruments, standardisation bases, user evidence, contexts of use, evidence integration, and identified gaps. This information was then grouped according to the knowledge functions required to advance the evaluation results towards the model formulation stage. The grouping followed the principle of traceability: each category was required to be traceable to one or more sources within the evidence mapping and was not permitted to introduce empirical claims that were not supported by the 54 studies. This process yielded the Design Knowledge Taxonomy as a conceptual artefact that bridges evidence mapping and model formulation.

IV. RESULTS AND DISCUSSION

4.1. Bibliometric Characteristics and Studies

Figure 2 shows that the number of publications between 2021 and 2025 fluctuated during the early years of the period before increasing sharply in the final two years. The number of publications decreased from approximately 9 articles in 2021 to 5 articles in 2022, before increasing to approximately 6 articles in 2023, 14 articles in 2024, and 20 articles in 2025. This trend indicates growing research attention to accessibility, usability, and learning technologies for deaf and hard-of-hearing (DHH) users, particularly since 2024.

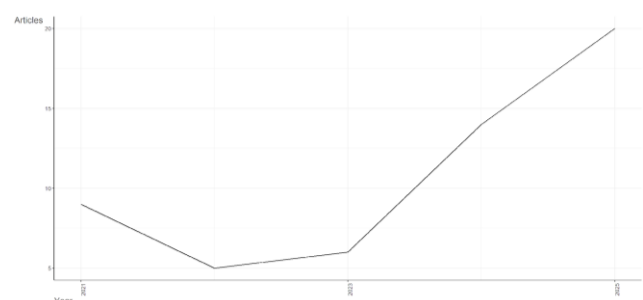


Fig. 2. Annual scientific output.

A bibliometric analysis of the 54 documents revealed that the publications were distributed across 38 sources (journals and proceedings) and involved 227 authors, with only one article authored by a single author. The average number of authors per document was 4.39, while the international collaboration rate was 31.48%. The annual publication growth rate was 22.09%. The corpus also contained 228 author keywords, with an average document age of 2.43 years and an average of 5.778 citations per document. Overall, these characteristics indicate that the corpus is relatively up to date and reflects an evolving research field characterised by a collaborative research environment.

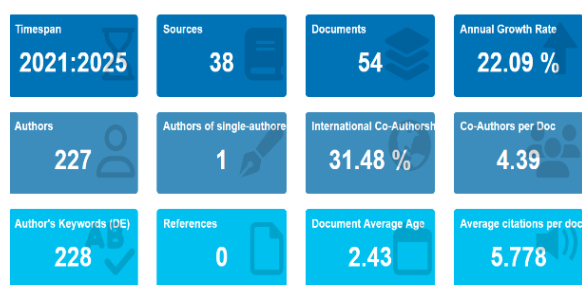


Fig. 3. Annual scientific output.

An analysis of keyword co-occurrence across the 54 studies yielded a network comprising 23 items distributed across four main clusters (Figure 4). The largest nodes were 'deep learning', 'sign language', and 'accessibility', indicating that these three themes occupy central positions within the research corpus.

The first cluster (green) is dominated by technical terms such as 'deep learning', 'machine learning', 'artificial intelligence', 'convolutional neural network', 'sign language recognition', 'gesture recognition', and variants of sign language, such as Indian Sign Language (ISL). It shows a clear emphasis on the development of algorithms and computational models for sign language and gesture recognition.

The second cluster (colored red) is concerned with accessibility and includes terms such as assistive technology, virtual reality, and deaf and hard of hearing (DHH) users. This cluster is a collection of papers which treat technology as an avenue towards greater access to education for DHH users.

The third cluster (colored blue) comprises more specific terms like Korean Sign Language (KSL), hand gesture recognition, angle features, and distance features. This theme reflects research focused on extracting hand movement features and developing models for specific sign languages.

The fourth cluster links the keyword 'learning' with several other themes, thereby serving as a bridge between the pedagogical and technical dimensions of the research.

The density of relationships among the keywords indicates the dominance of deep learning applications in sign language recognition and the development of assistive technologies. Conversely, the 'accessibility' node forms a relatively less dense cluster than the algorithmic clusters. This pattern suggests a gap between technological progress and the depth of 'quality of use' evaluation. Therefore, a more detailed mapping of the evaluation evidence across the 54 studies is required.

Figure 5 displays an overlay visualisation of the co-occurrence network based on the average publication year of the 54 studies. The colour scale indicates the chronology of theme emergence, ranging from blue and purple for earlier themes to green and yellow for more recent ones. The keyword 'accessibility' and related terms tend to appear in earlier publications, whereas 'deep learning', 'convolutional neural networks', 'gesture recognition', 'real-time translation', and specific sign languages feature more prominently in recent publications.

The overlay pattern reveals a shift from more general studies of accessibility towards the increasingly intensive use of deep learning and sign language recognition. Nevertheless, 'accessibility' and 'learning' remain important connecting nodes within the network. These findings underscore the need to assess whether the growth in technical innovation is accompanied by sufficient evidence of usability, accessibility, user engagement, and learning outcomes.

The thematic map in Figure 6 shows that deep learning-based sign language recognition and assistive technology are the main driving themes, whereas accessibility and e-learning for deaf and hard-of-hearing (DHH) users remain relatively underdeveloped. These findings further demonstrate that technical innovation is advancing faster than research that explicitly evaluates the quality of use and accessibility evidence in learning contexts.

4.2. Methods and Evaluation Metrics

The characteristics of the 54 studies included the distribution of publication years, types of systems or contributions, platforms, application contexts, and reporting on DHH participants. A summary of these characteristics was used to provide a context for comparing the evaluation methods and metrics reported across the studies. The classification of systems, platforms, and contexts was based on information available in the titles, abstracts, and full texts. Information that was not reported was coded as NR.

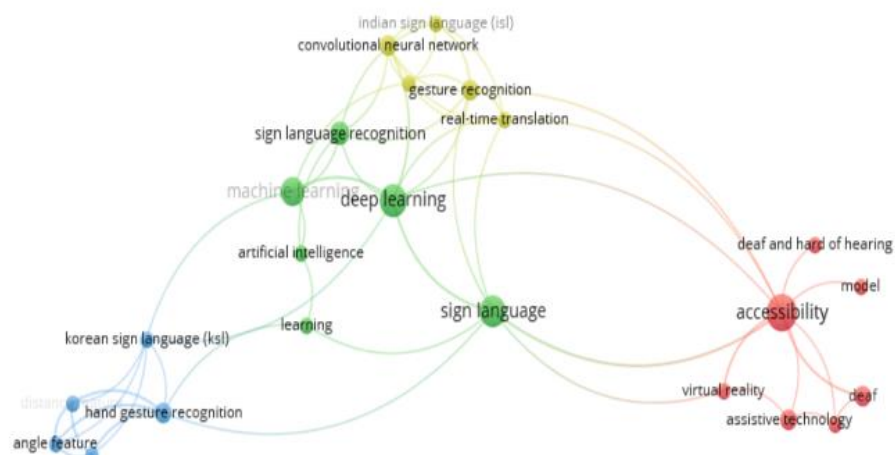


Fig. 4. Keyword relationship map (cluster visualisation).

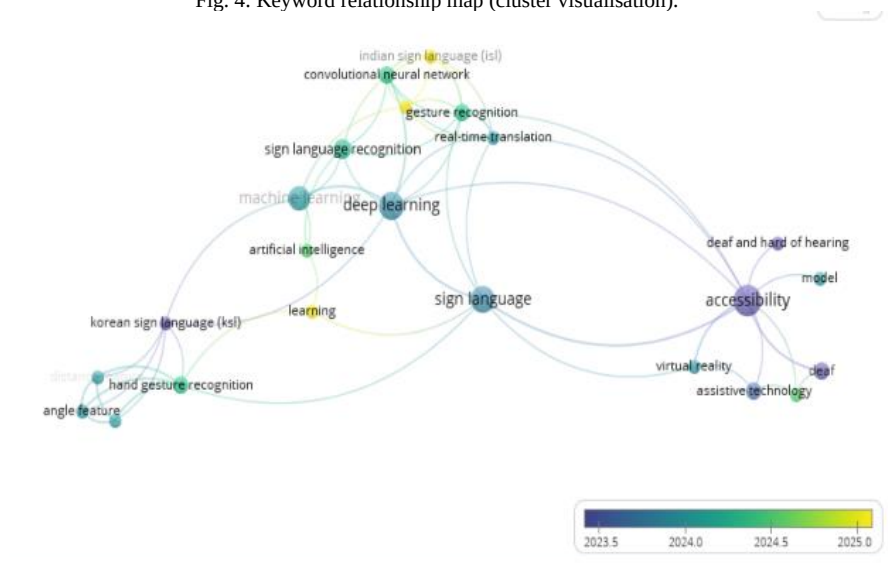


Fig. 5. Overlay visualization.

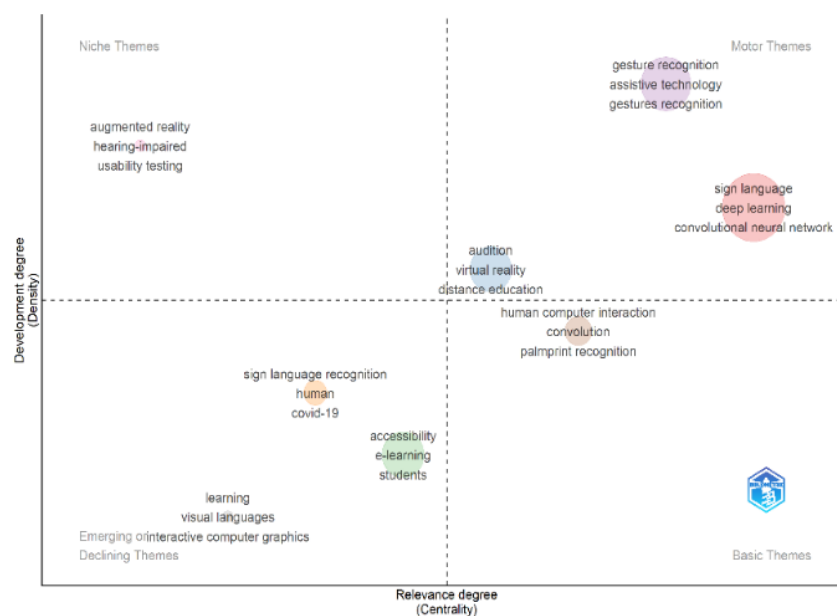


Fig. 6. Thematic map.

Table 1 shows that publications were concentrated in 2024–2025. In terms of system type, the VR/AR/XR category was relatively dominant, followed by ML/CV-based sign language recognition and translation models. Many studies were conducted in educational or training contexts without specifying a particular educational level. Reporting on DHH participants was also inconsistent: 13 studies explicitly reported the involvement of DHH participants, 22 studies did not involve DHH participants or were not relevant to this category, and 19 studies did not report participant information (NR). Once the study characteristics had been mapped, the analysis proceeded by comparing the evaluation methods with the aspects being assessed.

Table 2 shows that technical performance metrics were the most frequently used evaluation approach, appearing in 28 studies. User-based methods were used to a more limited

extent: the System Usability Scale (SUS) was used in five studies, the USE Questionnaire in one study, specialised surveys/Likert scales in nine studies, interviews in four studies, observations in three studies, and heuristic/expert reviews in four studies. All frequencies were calculated based on the accessible full texts.

Regarding accessibility, formal audits based on WCAG or Section 508 were reported in only one study, whereas specific evaluations of captions and subtitles were reported in three studies. This pattern suggests that accessibility was more frequently represented by the presence of accessibility features than verified through documented compliance audits. Measurements of learning outcomes using pre–post tests were also reported in only four studies. Overall, evidence of technical performance was far more prevalent than evidence of standardised accessibility, user-centred usability, and learning effectiveness.

Table 1. Summary of the characteristics of the 54 studies.

Aspect	Category	N (Number of Articles)	%
Year of publication	2025	20	37
Year of publication	2024	14	25.9
Year of publication	2023	6	11.1
Year of publication	2022	5	9.3
Year of publication	2021	9	16.7
Type of system/contribution	VR/AR/XR learning/experience system	23	42.6
Type of system/contribution	Sign language recognition/translation models (ML/CV)	12	22.2
Type of system/contribution	Accessibility of captions/subtitles & MOOC content	7	13.0
Type of system/contribution	Wearable assistive devices	6	11.1
Type of system/contribution	UX/accessibility studies (qualitative/mixed-methods)	4	7.4
Type of system/contribution	Learning platform / VLE	2	3.7
Platform	VR/AR/XR (headset/immersive)	24	44.4
Platform	Unspecified/algorithmic	17	31.5
Platform	Web/LMS	7	13.0
Platform	Wearable	6	11.1
Context	Education/training (unspecified level)	31	57.4
Context	Online learning/distance learning	9	16.7
Context	Other	8	14.8
Context	Library/orientation	2	3.7
Context	Schools (K-12/special schools)	1	1.9
Context	Museums/heritage	1	1.9
Context	Higher education	1	1.9
Context	Auditorium/performance experience	1	1.9
Reporting by participants with hearing and speech impairments	No participants with hearing and speech impairments / not relevant (0)	22	40.7
Reporting by participants with hearing and speech impairments	NR (not reported)	19	35.2
Reporting by participants with hearing and speech impairments	DHH participants reported (>0)	13	24.1

Table 2. Mapping of evaluation methods against the aspects assessed (RQ1).

Method/instrument	Key dimensions	n	Use in studies
Technical performance metrics	Technical	28	Accuracy, precision, recall, F1 score, mAP, latency, and error rate; generally, without user testing.
SUS (System Usability Scale)	Usability, UX	5	SUS score (0–100) to assess the usability of a system or application.
USE Questionnaire	Usability, UX	1	Usefulness, satisfaction, ease of use, and ease of learning.
Specialised survey / Likert scale	Usability, Accessibility, UX	9	Satisfaction/acceptance; items are often non-standardised and vary.
Interview	Usability, Accessibility, UX	4	Exploring barriers, needs, experiences and usage contexts.
Observation	Usability, Accessibility, UX	3	Observation of tasks/activities; can be combined with task metrics.
Heuristic/expert review	Usability, Accessibility	4	Expert evaluation; quick but requires clear reporting procedures.
Pre–post learning test	Learning outcomes	4	Measures the impact of media/software on learning outcomes.
Accessibility audit (WCAG/508)	Accessibility	1	Checklist/compliance with standards; ideally supplemented by user testing.
Specific evaluation of subtitles/captions	Accessibility, Learning Outcomes, Technical	3	Quality of captions/subtitles: synchronisation, positioning, legibility and user ratings.

The quality assessment of the 54 studies yielded a mean score of 5.93, a median score of 6, and a range of 3–8. A total of 63.0% of the studies were classified as high quality, while 37.0% were classified as moderate quality. The quality assessment was used to support the interpretation of the results rather than as an exclusion criterion. Although methodological reporting across the corpus was generally adequate, the comprehensiveness of the reported evidence types remained uneven, particularly with regard to formal accessibility audits, the involvement of DHH participants, learning outcomes, and contexts of use.

4.3. Evaluation Evidence Taxonomy

The results of the data extraction and evaluation mapping were then organised into six dimensions of the Evaluation Evidence Taxonomy, enabling diverse methods and forms of evidence to be compared within a consistent framework. The taxonomy retains the information reported in the 54 studies and serves as a classification structure for organising evaluation evidence. It does not prescribe the sequence of the QA process or define any mandatory model components; rather, its function is limited to grouping the forms of evidence reported across the corpus.

- (i) Evaluation targets: technical performance, accessibility, usability/UX, and learning outcomes.
- (ii) Methods/instruments: computational metrics, SUS/USE/Likert scales, interviews, observations, expert reviews, and pre–post tests.
- (iii) Basis for standardisation: formal standards, standardised instruments, or custom/non-standard instruments, including WCAG/Section 508, SUS, USE, custom surveys, and caption evaluations.
- (iv) User evidence: DHH participants involved, no participants/not applicable, or NR. The corpus contains 13 studies involving DHH participants, 22 studies with no participants or where participant involvement was not applicable, and 19 studies in which participant information was not reported (NR).
- (v) Context of use: reported context or NR, covering educational settings, platforms, devices, and conditions of use.
- (vi) Evidence integration: technical-only, user-based, or mixed evidence. The classification comprises EVAL-

TECH (23 studies), EVAL-USER (8 studies), and EVAL-MIXED (21 studies).

4.4. Evidence-Gap Map

An evidence-gap map was compiled to highlight areas of evaluation that remain under-represented across the 54 studies. The growth in publications between 2024 and 2025 and the dominance of deep learning and sign language recognition indicate a high level of technical innovation. However, the evaluation mapping indicates that these developments have not always been accompanied by a commensurate level of evidence regarding quality of use. Key gaps include the dominance of technical performance metrics, limited involvement of DHH participants, the scarcity of formal accessibility audits, limited measurement of learning outcomes, and inconsistent reporting of contexts of use. This pattern is consistent with recent literature emphasising that accessibility should be assessed in relation to usability, comprehensibility, and user experience in real-world contexts (Ahn et al., 2025; Loaiza-Ramírez and González-beltrán, 2026; Pérez-Arteaga et al., 2025; Sivunen, 2025).

The evidence-gap map confirms that high algorithmic performance cannot be directly equated with high levels of accessibility or usability. To ensure that the interpretation of the results is transparent and facilitates comparison across studies, technical evidence needs to be distinguished from user-based evidence, standards-based evidence, and outcome evidence.

The six dimensions of the Evaluation Evidence Taxonomy provide a framework for comparing studies based on the evidence reported in each study. Together with the evidence-gap map, this taxonomy forms the basis for the design-oriented synthesis conducted by the researchers. Thus, design knowledge is derived from evidence that has been systematically analysed and classified rather than from generalisations drawn from outside the corpus.

The identified gaps also highlight several reporting priorities for future research, particularly the characteristics of DHH participants, contexts of use, bases for standardisation, relationships between methods and evidence, and the measurement of learning outcomes. Table

3 summarises the available evidence, its implications for interpretation, and the identified reporting priorities.

At this stage, the main outputs of the evidence mapping are the Evaluation Evidence Taxonomy and the evidence-gap map. Both outputs remain evidence-based and descriptive in nature and are not intended to constitute a quality assurance (QA) framework or prescribe specific procedures. These two outputs then provide the basis for the researchers to derive the Design Knowledge Taxonomy through design-oriented synthesis.

4.5. Design Knowledge Taxonomy

Design-oriented synthesis of the Evaluation Evidence Taxonomy and the evidence-gap map yielded the knowledge clusters required to advance the evaluation evidence towards the model formulation stage. This process was conducted by

the researchers while maintaining traceability between the categories developed and their underlying sources of evidence. The relationships considered primarily encompassed the links between user evidence and the context of use, the relationship between the basis for standardisation and evaluation methods, and the need to connect evidence integration with traceability and improvement priorities.

The results of this synthesis formed the five categories of the Design Knowledge Taxonomy: (1) User and Context Needs, (2) Quality Needs, (3) Evaluation Methods, (4) Evidence and Traceability, and (5) Improvement and Research Priorities. These five categories are not treated as elements of the final QA Model but rather as knowledge clusters that need to be established as the research progresses from evidence mapping towards model formulation.

Table 3. Evidence-gap map for the evaluation of learning systems for people with hearing impairments.

Evidence gaps	Evidence from 54 studies	Implications for interpretation	Reporting priorities
Technical evaluations predominate over user-based evaluations	EVAL-TECH = 23; EVAL-USER = 8; EVAL-MIXED = 21. Technical metrics were used in 28 studies.	Algorithmic performance may appear high even though ease of use and the experience of people with hearing and speech impairments have not been demonstrated.	Separate technical evidence from usage evidence; report on user testing and the limits of generalisation.
Reporting on participants with hearing and speech impairments is inconsistent	DHH >0 = 13/54; none/not applicable = 22/54; NR = 19/54.	The relevance of the findings to the DHH population and cross-study comparisons are difficult to assess.	Report the number/criteria of participants, communication/language characteristics, technological experience, and limitations.
Formal accessibility audits are very rare	WCAG/Section 508 audits were found in approximately 1 out of 54 studies.	The presence of captions or multimodal features may be considered as accessibility without evidence of compliance.	Report standards, audit procedures, criteria, and results; do not equate the presence of features with compliance.
Evidence of learning outcomes is limited	Pre-post tests or explicit learning outcomes were found in around 4 out of 54 studies.	Claims that media support learning are not always supported by evidence of outcomes.	Distinguish claims regarding technical performance from learning effectiveness and use outcome measures appropriate to the objectives.
Reporting on the real-world context of use is often incomplete	The context of use was coded as 'NR' in 10 out of 54 studies.	Replicating and transferring results to different schools, devices, or support conditions is difficult.	Report the setting, device, duration, support/facilitator, and relevant conditions of use.

Table 4. Design Knowledge Taxonomy derived from the researchers' synthesis of evidence.

Design knowledge categories	Sources in the evidence mapping	Functions for the model formulation stage
User knowledge and context	User evidence; context of use; reports from participants with hearing and speech impairments	Defining user boundaries, communication characteristics, settings, devices and conditions of use that must be considered.
Knowledge of quality requirements	Evaluation targets; standards; gaps in accessibility, usability and learning outcomes	Formulating the quality requirements that need to be demonstrated, without turning them into a universal list of features.
Knowledge of evaluation methods	Technical metrics; SUS/USE/Likert; interviews; observation; audits; pre-post	Selecting a combination of methods appropriate to the evidence targets and avoiding an over-reliance on purely technical evaluation.
Knowledge of evidence and traceability	Evidence integration; reporting completeness; basis for standardisation	Determining the minimum evidence required and the relationship between claims, methods, results and the basis for decision-making.
Knowledge of improvements and research priorities	Evidence-gap map	Identifying areas requiring strengthening: DHH involvement, formal accessibility audit, learning outcomes, and context-of-use reporting.

The Design Knowledge Taxonomy maintains the link between evaluation findings and design needs without prematurely defining the structure of the model. 'User and

Context Needs' summarises the information requirements concerning DHH users and the conditions of use; 'Quality Needs' links quality targets to relevant standards and identified gaps; 'Evaluation Methods' summarises

methodological choices in relation to the targeted evidence; 'Evidence and Traceability' emphasises the interrelationships among claims, methods, results, and the basis for decision-making; and 'Improvement and Research Priorities' identifies areas that require further strengthening in future research.

The Design Knowledge Taxonomy is an intermediate research artefact that provides a structured body of knowledge to support the subsequent QA Model formulation stage. This taxonomy does not establish causal relationships, assign weights, define mandatory sequences, determine readiness scores, or prescribe the structure of the final model. This distinction is important to ensure that the research output is understood as a bridge between evidence mapping and model formulation rather than as a completed and empirically validated model.

4.6. Discussion: From Evaluation Evidence to Design Knowledge

The key findings indicate a gap between the pace of technical innovation and the comprehensiveness of the available evidence on quality of use. The predominance of technical metrics across 28 studies, compared with only one formal accessibility audit and four measurements of learning outcomes, suggests that technological performance alone is insufficient to conclude that a system is truly accessible, usable, or effective for learning. The Evaluation Evidence Taxonomy and the evidence-gap map make this disparity more apparent, while the Design Knowledge Taxonomy organises these findings into design-oriented knowledge that remains traceable to the underlying evidence.

These findings are consistent with recent literature indicating that accessibility for DHH users cannot be adequately assessed based solely on the presence of accessibility features. Loaiza-Ramírez and González-Beltrán (2026) position the specific needs of DHH users as a basis for developing digital accessibility guidelines, while Ahn et al. (2025) and Pérez-Arteaga et al. (2025) demonstrate that the quality of caption presentation should be assessed in terms of how information is represented and received by users. The literature also highlights the importance of DHH users' preferences in the evaluation of captions. Overall, these studies reinforce the findings of this review that the mere presence of accessibility features cannot be treated as evidence of quality without evaluating their actual use, comprehensibility, and user experience.

The limitations in reporting participant characteristics and contexts of use identified in this review also have important implications. Sivunen (2025) demonstrates that accessibility experiences in technology-enhanced learning are linked to language practices and usage experiences within the deaf and hard-of-hearing community. In the context of the findings of this study, this reinforces the need to consider user evidence together with context-of-use evidence when evaluating the quality of learning systems. Reporting on settings, devices, forms of support, and users' communication characteristics is essential to ensure that evaluation results can be interpreted and compared accurately.

The two taxonomies serve different but complementary functions. The Evaluation Evidence Taxonomy describes how evaluation evidence is reported in the literature, while the Design Knowledge Taxonomy organises the types of knowledge required to translate that evidence into the design phase. This relationship extends the findings of the systematic literature review (SLR) mapping without automatically converting the evidence into model components. The transition from evaluation evidence to design knowledge therefore takes place through the researchers' analysis and consolidation while maintaining a distinction between empirical findings, interpretations, and design requirements. In this way, the contribution of this study extends beyond identifying evidence gaps by providing a traceable structure for bringing evaluation results into the subsequent QA Model development stage.

4.7. Conceptual Directions for the Development of the Next QA Model

Based on the Evaluation Evidence Taxonomy, the evidence-gap map, and the Design Knowledge Taxonomy, the findings of this research can be summarised into four conceptual areas relevant to the next stage of QA Model development: User & Context, Quality Requirements, Evaluation & Evidence, and Improvement. These four areas are not components of the final QA Model but rather conceptual directions that summarise the types of knowledge that need to be connected when the model is subsequently formulated. Traceability is positioned as a cross-cutting principle to ensure that requirements, criteria, evaluation methods, evidence, decisions, and recommendations for improvement can be traced back to their underlying sources.

Figure 7 illustrates these conceptual directions at a deliberately generalised level. 'User & Context' provides the basis for understanding the characteristics of DHH users and their conditions of use. 'Quality Requirements' translates these needs into objectives related to accessibility, usability, and learning support. 'Evaluation & Evidence' connects quality requirements with methods, instruments, user-based evidence, standards-based evidence, and learning outcomes. Meanwhile, 'Improvement' maintains the relationship between findings, priorities, recommendations, and re-evaluation. This representation does not specify a mandatory sequence, weighting scheme, decision thresholds, or the architecture of the final model.

In the subsequent stage of QA Model development, these four conceptual areas can be used in conjunction with the Evaluation Evidence Taxonomy, the evidence-gap map, and the Design Knowledge Taxonomy as a structured foundation for the model formulation process. This stage falls outside the scope of RQ1–RQ3 addressed in this study. The relationships among these areas, alternative structures, evaluation mechanisms, forms of evidence, and refinement processes will be formulated during the subsequent model development stage. As this stage has not yet been undertaken, this study does not specify the architecture of the final model. The planned Human–AI Co-Creation mechanism to support this process is presented separately in Figure 8.

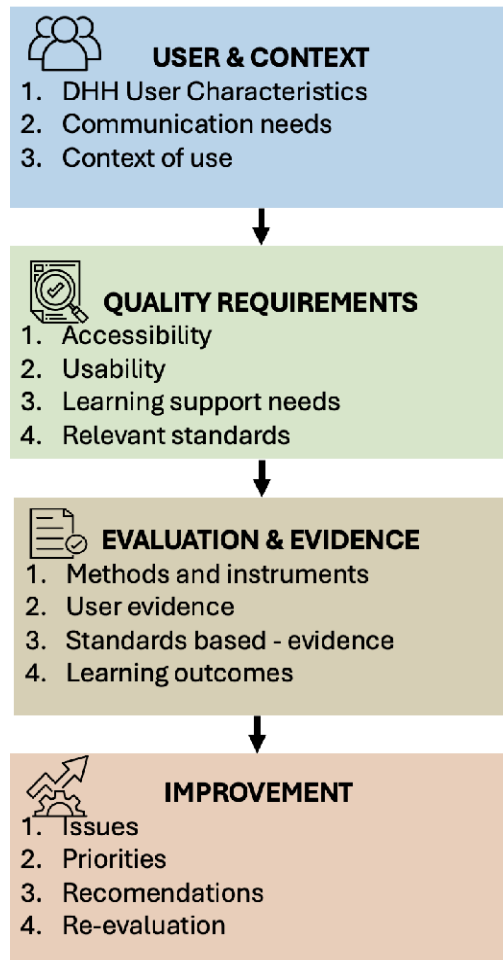


Fig. 7. Conceptual direction for the development of the next QA Model derived from the synthesis of evidence.

4.8. Proposed Human–AI Co-Creation for Further QA Model Development

To clarify the role of Generative AI in relation to the model to be developed, Figure 8 presents a proposed conceptual mechanism for the subsequent stages of QA Model development. This mechanism has not yet been implemented in the present study. The initial inputs are derived from outputs formulated by the researchers, namely the Evaluation Evidence Taxonomy, the evidence-gap map, the Design Knowledge Taxonomy, and relevant DHH standards and contexts. The LLM is intended to function as a design-support tool to assist in mapping relationships, exploring and comparing alternative structures, and checking conceptual consistency. Thus, the LLM is not a component of the QA Model structure; rather, its role lies within the model development process as a design-support mechanism that remains under human control.

This mechanism clearly distinguishes the generation of research evidence from AI-supported activities during the design stage. The researcher remains the ‘Human-in-the-Loop’, interpreting the LLM’s outputs, checking their traceability to the source evidence, selecting, correcting, accepting, or rejecting alternatives, and carrying out further refinements. The LLM can then be used iteratively to reformulate alternatives based on the researcher’s feedback.

Expert validation is positioned as a planned human-led stage for assessing the feasibility and suitability of candidate models. Thus, the expected output of the model development stage is a set of alternative QA Model designs that are ready for further validation, rather than a final model determined by the LLM (Clark et al., 2025; Khraisha et al., 2024; Shah, 2025).

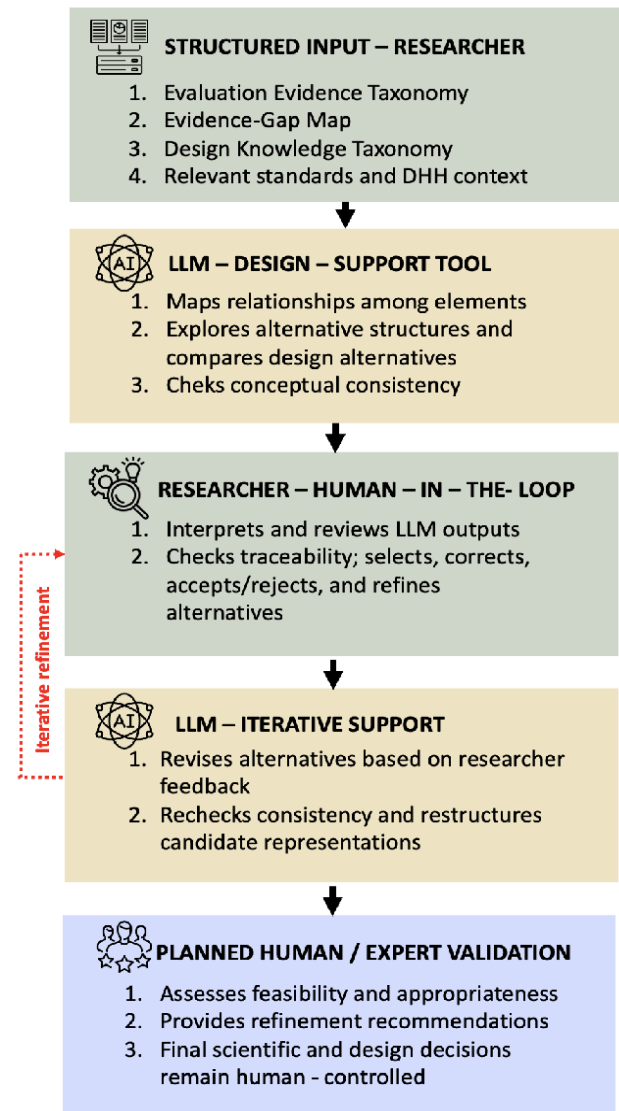


Fig. 8. Proposed Human–AI Co-Creation mechanism for the subsequent QA Model development stage.

4.9. Implications for Research and Practice

The first methodological implication is the need to explicitly distinguish between evidence of technological performance and evidence of use. Research on DHH learning systems should report evaluation targets, methods, standardisation frameworks, user characteristics, contexts of use, and types of evidence separately before integrating them. The presence of features such as captions or visual support cannot be directly regarded as evidence of compliance or effectiveness; such claims must be supported by standards-based evidence, user-based evidence, and outcome evidence that are aligned with the evaluation

objectives. This approach is consistent with recent studies emphasising the importance of the specific needs of DHH users, caption quality, and user experience in digital accessibility assessments (Ahn et al., 2025; Loaiza-Ramírez and González-beltrán, 2026; Pérez-Arteaga et al., 2025).

The second implication relates to the development of QA Models and developer practices. The Evaluation Evidence Taxonomy helps identify the types of evidence available, while the Design Knowledge Taxonomy summarises the knowledge required prior to model design. The conceptual framework in Figure 7 consolidates these outputs into four areas that can serve as a starting point for model formulation, while Figure 8 illustrates the planned Human–AI mechanism for exploring and refining alternatives during subsequent stages of QA Model development. Both representations remain conceptual in nature. Decisions regarding model structure, inter-component relationships, priorities, and feasibility must be established through a documented design process, researcher assessment, and expert validation rather than simply being derived from patterns in the literature or LLM outputs. Consequently, the practical implication of this research is the provision of an explicit evidence structure that enables developers and evaluators to trace the relationships among user needs, standards foundations, evaluation methods, generated evidence, and recommendations for improvement before design decisions are finalised.

4.10. Limitations

This study has several limitations. First, the corpus is restricted to the Scopus database and the period from 2021 to 2025; consequently, it does not represent the entire body of potentially relevant literature. Second, variations in reporting across the primary studies limit the depth and consistency of the classification, particularly with regard to DHH participant characteristics and contexts of use. Third, the Design Knowledge Taxonomy and the conceptual framework presented in Figure 7 were derived through the researchers' synthesis of the evidence mapping and have not yet been validated as a QA Model through expert judgement or implementation in specific software environments. Fourth, the Human–AI Co-Creation mechanism presented in Figure 8 remains a proposed mechanism for the subsequent QA Model development stage and has not yet been implemented. Consequently, its effectiveness, reproducibility, and contribution to model design quality cannot yet be determined based on the findings of this study.

V. CONCLUSION

This study examined 54 studies on learning systems or media for deaf and hard-of-hearing (DHH) users published between 2021 and 2025. The evaluation mapping revealed a predominance of technical performance metrics, while user-based evaluations, formal accessibility audits, the involvement of DHH participants, reporting of contexts of use, and learning-outcome measures remained limited. Based on the researchers' analysis, this study produced a six-dimensional Evaluation Evidence Taxonomy and an evidence-gap map. These two outputs were subsequently synthesised into a Design Knowledge Taxonomy comprising User and Context Needs, Quality Needs, Evaluation

Methods, Evidence and Traceability, and Improvement and Research Priorities. To bridge these findings towards the model formulation stage, this study further consolidated the conceptual direction into four areas: User & Context, Quality Requirements, Evaluation & Evidence, and Improvement, with traceability positioned as a cross-cutting principle. This representation constitutes an intermediate research artefact rather than a final, validated QA Model. In the subsequent stage of QA Model development, this artefact can serve as structured input for LLM-based Human–AI Co-Creation under a Human-in-the-Loop approach, as illustrated by the conceptual mechanism in Figure 8. The LLM is positioned outside the QA Model structure but within the model development process as a design-support tool, while interpretation, selection, refinement, validation, and final decision-making remain under human control. This mechanism has not yet been implemented and is therefore not presented as a research result of this study.

ACKNOWLEDGMENT

We acknowledge Universitas Komputer Indonesia.

REFERENCES

- Abbood, A. M. H. (2024). Effectiveness of cooperative learning using multimedia in some physical abilities and basic skills for junior players in basketball. *ASEAN Journal of Physical Education and Sport Science*, 3(1), 7–16.
- Ahn, S., Kim, J., Shin, C., & Hong, J.-H. (2025). Visualizing speech styles in captions for deaf and hard-of-hearing viewers. *International Journal of Human-Computer Studies*, 194, 103386.
- Al Huseini, D. N., & Rahmat, M. (2023). Computational bibliometric analysis: Can digital transformation improve the quality of Islamic learning? *ASEAN Journal of Religion, Education, and Society*, 2(2), 123–138.
- Al-Saidi, M., Ballagi, Á., Hassen, O. A., & Saad, S. M. (2024). Type-2 neutrosophic Markov chain model for subject-independent sign language recognition: A new uncertainty-aware soft sensor paradigm. *Sensors*, 24(23), 7828.
- Ariya, P., Yensathit, Y., Thongthip, P., Intawong, K., & Puritat, K. (2024). Assisting hearing and physically impaired students in navigating immersive virtual reality for library orientation. *Technologies*, 13(1), 2.
- Ariyanti, L. S., & Nandiyanto, A. B. D. (2022). Distance teaching of stress materials to junior high school students using digital media. *ASEAN Journal of Community Service and Education*, 1(1), 7–14.
- Babalian, V., & Slutu, S. (2025). Real-time sign language-to-speech translation system with AI-powered wearable technology. *IEEE Access*, 13(September), 151652–151663.
- Berrezueta-Guzman, S., Daya, R., & Wagner, S. (2025). Virtual reality in sign language education: opportunities, challenges, and the road ahead. *Frontiers in Virtual Reality*, 6, 1625910.

- Cabanillas-Carbonell, M., Cusi-Ruiz, P., & Prudencio-Galvez, D. (2022). Mobile application with augmented reality to improve the process of learning sign language. *IJIM*, 16(11), 51–64.
- Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., & O'Neill, H. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*, 16, 601–619.
- Coster, M. De, Shterionov, D., Herreweghe, M. Van, & Dambre, J. (2024). Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23(3), 1305–1331.
- Escudeiro, P., Galasso, B., Teixeira, D., Gouveia, M. C., & Escudeiro, N. (2022). Inclusive MOOC--educational content for deaf people, a Portuguese proof of concept. *Cypriot Journal of Educational Sciences*, 17(4), 1285–1303.
- González-Rodríguez, J. R., Córdova-Esparza, D. M., Terven, J., & Romero-González, J. A. (2024). Towards a bidirectional Mexican sign language–Spanish translation system: A deep learning approach. *Technologies*, 12(1), 7.
- Hashi, A. O., Mohd Hashim, S. Z., Mirjalili, S., KEBANDE, V. R., Al-Dhaqm, A., Nasser, M., & A Samah, A. B. (2025). A hybrid CNN-transformer framework optimized by Grey Wolf algorithm for accurate sign language recognition. *Scientific Reports*, 15(1), 43550.
- Heck, G. S., & Ferraro, J. L. (2025). Accessible practices in science museums and their impact on the science capital of visually and hearing-impaired people. *Rev. Int. de Pesquisa em Didática das Ciências e Matemática (RevIn)*, 6, e025011.
- Khraisha, Q., Put, S., Kappenberg, J., Warratich, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4), 616–626.
- Kim, N., Choe, H., Lee, S., & Kang, C. (2025). Design and implementation of a 3D Korean sign language learning system using pseudo-hologram. *Applied Sciences*, 15(16), 8962.
- Li, J., Xu, J., Liu, Y., Xu, W., & Li, Z. (2024). Enhancing the applicability of sign language translation. *IEEE Transactions on Mobile Computing*, 23(9), 8634–8648.
- Loaiza-Ramírez, I., & González-Beltrán, B. A. (2026). Digital accessibility guidelines for deaf and hard of hearing persons: A state-of-the-art review article. *Universal Access in the Information Society*, 25(1), 4.
- Lynch, S., Brunton, J., & Farrell, O. (2026). Experiences of disabled students in online education: A systematic review. *Distance Education*, 47(1), 70–90.
- Millett, P. (2021). Accuracy of speech-to-text captioning for students who are deaf or hard of hearing. *Journal of Educational, Pediatric & (Re)Habilitation Audiology*, 25, 1.
- Musayaro, S., Asmiati, N., Utami, Y. T., Mulia, D., Sidik, S. A., Abadi, R. F., Pratama, T. Y., Maslahah, S., & Pramudyo, A. S. (2023). A digital accessibility and inclusive design-based e-module in higher education: Does it work in a classroom with a deaf student? *ASEAN Journal of Community and Special Needs Education*, 2(1), 55–60.
- Nandiyanto, A. B. D., & Yulistiani, F. (2024). How technical language supports students' understanding in engineering education: Definitions and step-by-step interpretation of dimensionless parameters (Re, Pr, Pe, Nu, Gr, Ra, Bi, and Fo) in heat transfer. *International Journal of Language Education*, 8(4), 885–900.
- Nedjar, I., & Mohammed, M. (2024). Interactive system based on artificial intelligence and robotic arm to enhance Arabic sign language learning in deaf children. *Education and Information Technologies*, 29, 24563–24580.
- Ogundele, A. G., Umar, I. Y., & Idris, A. M. (2025). The emergence of new technologies in metalwork/automobile industries: Issues, challenges and opportunities in emanating from for delivery of technical education on a pandemic era. *ASEAN Journal for Science Education*, 4(1), 1–6.
- Panopoulos, I., Topalis, E., Petrellis, N., & Hadellis, L. (2025). Greek sign language detection with artificial intelligence. *Electronics*, 14(16), 3241.
- Perea-Trigo, M., Vázquez-Enríquez, M., Benjumea-Bellot, J. C., Alba-Castro, J. L., & Álvarez-García, J. A. (2025). Sign language anonymization: Face swapping versus avatars. *Electronics*, 14(12), 2360.
- Pérez-Arteaga, J., Letón, E., Pérez Martín, J., Molanes-López, E., & Rodríguez-Ascaso, A. (2025). Assessment of captions' quality by students of a MOOC on accessibility of digital materials. *Computers and Education Open*, 9, 100279.
- Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, 164, 113794.
- Reyes, J. I., Meneses, J., & Xavier, M. (2023). Suitability of online higher education for learners with disabilities: The students' voices. *Journal of Special Education Technology*, 38(3), 370–383.
- Shah, C. (2025). From prompt engineering to prompt science with humans in the loop. *Communications of the ACM*, 68(6), 54–61.
- Shahin, N., & Ismail, L. (2024). From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation. *Artificial Intelligence Review*, 57(10), 271.

- Shin, J., Musa Miah, A. S., Hasan, M. A. M., Hirooka, K., Suzuki, K., Lee, H. S., & Jang, S. W. (2023). Korean sign language recognition using transformer-based deep neural network. *Applied Sciences*, 13(5), 3029.
- Sivunen, U. (2025). Lived experiences of language use and (in) accessibility among multilingual deaf and hard-of-hearing adolescents and young people in technology-enhanced learning in Finland. *Deafness & Education International*, 27(4), 340-366.
- Susilawati, A. (2024). A bibliometric analysis of global trends in engineering education research. *ASEAN Journal of Educational Research and Technology*, 3(1), 103–110.
- Tan, S., Khan, N., An, Z., Ando, Y., Kawakami, R., & Nakadai, K. (2024). A review of deep learning-based approaches to sign language processing. *Advanced Robotics*, 38(23), 1649–1667.
- Violet, I. M., & Sri, R. L. (2025). A comprehensive survey on recent advances and challenges in sign language recognition systems. *Discov. Artif. Intell.*, 5(1), 419.