Explainable and Context-Aware Speech-to-Text for Special and Inclusive Education: A Systematic Review and a Proposed Conceptual Framework

Sabnam Kumari[1], Anu Saini[2], Sunita Kumari[3], Deepak[4], Vaani Garg[5] Bhawna Bhardwaj[6*]

[1,4,5,6]Department of Computer Science & Engg, Maharaja Surajmal Institute of Technology, New Delhi, India

Shabnam022@gmail.com, deepak@msit.in, vaani.garg1708@gmail.com

[2,3]Department of Computer Science & Engg, G.B Pant, DSEU Campus, Okhla, New Delhi, India

anuanu16@gmail.com, sunita2009@gmail.com,

*Corresponding author: Bhawna Bhardwaj, bhawnabhardwaj@msit.in

HOW TO CITE:

Sabnam Kumari, Anu Saini, Sunita Kumari, Deepak, Vaani Garg, Bhawna Bhardwaj (2026). Explainable and Context-Aware Speech-to-Text for Special and Inclusive Education: A Systematic Review and a Proposed Conceptual Framework. International Journal of Special Education, 41(9s), 589-607

ABSTRACT:

Background: Speech-to-text (STT) systems built on automatic speech recognition (ASR) are widely used to caption lectures and classroom talk and to support learners who are deaf or hard of hearing, who have communication or language-processing differences, or who learn in multilingual settings. Recognition accuracy has improved markedly, yet errors persist in authentic conditions and are typically delivered without any indication of which words are unreliable or why.

Objective: This review synthesises evidence on the reliability, transparency, and explainability of STT used as an access technology in special and inclusive education, and identifies the gap that motivates a context-aware, explainable post-recognition approach.

Method: Following the PRISMA 2020 statement, peer-reviewed, English-language records (January 2009–May 2026) were sought in Scopus, Web of Science, ERIC, IEEE Xplore, and PubMed, with citation and hand searching. Records were screened against pre-specified eligibility criteria; included studies were charted and synthesised narratively across four themes.

Result: The synthesised evidence shows that (a) STT meaningfully improves access for learners with disabilities but its usefulness is bounded by accuracy and reliability; (b) recognition degrades for atypical speech, child speech, and non-mainstream dialects, raising equity concerns; (c) confidence and post-recognition correction can flag errors but are rarely exposed to users; and (d) explainability—though central to trustworthy educational AI—has scarcely been applied to STT. No identified system combined reliability awareness with user-facing explanation in an accessibility context.

Conclusions: The review reveals a transparency gap and derives a conceptual framework, ECASTT (Explainable Context-Aware Speech-to-Text Technology), that augments any ASR engine with multi-signal reliability assessment and human-readable explanation. The framework is presented as a research agenda, not an evaluated system, and is aligned with Universal Design for

Learning. Implications for assistive-technology design and a programme of user-centred validation are outlined..

Keywords: Assistive Technology; Speech-to-text; Automatic Speech Recognition; Explainable Artificial Intelligence; Accessibility; Explainable Context-Aware.

1. Introduction

Inclusive education commits education systems to equitable access to instructional content and to participation for all learners, including those with disabilities and diverse communication profiles (UNESCO, 2020). Digital accessibility technologies are central to that commitment, and speech-to-text (STT) systems occupy a particularly important place among them. By converting spoken language into a textual representation, STT enables learners to access lectures, classroom discussion, and recorded material through reading, replay, and search, and it underpins live captioning for learners who are deaf or hard of hearing (DHH) (Kawas et al., 2016; Wald, 2006).

The technical foundations of STT have advanced rapidly. Early systems combined hidden Markov models with handcrafted features (Rabiner, 1989); deep neural acoustic models (Hinton et al., 2012) and end-to-end architectures based on connectionist temporal classification (Graves et al., 2006) and attention (Chan et al., 2016) simplified the pipeline and improved accuracy; and self-supervised and large-scale weakly supervised models such as wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) produced robust, multilingual recognition. As per the standard benchmarks (Baevski et al., 2020), the big self-supervised models are claiming word error rates lower than 3%.

Although we have made progress on both fronts, there are still two problems that accuracy in and of itself does not resolve, and both matter for learners who rely on transcripts as their main access channel. Typically, ASR systems output the most probable transcript of the audio as well as little indication of which words are most reliable. An expression can have a high acoustic confidence but be wrong or inconsistent in the context and domain.. Second, errors are frequently invisible to the user; when an incorrect word is fluent and grammatical, a learner may absorb distorted instructional content with no signal that review is warranted. These limitations are compounded by equity concerns, because recognition is known to degrade for atypical speech, child speech, and non-mainstream dialects (Caballero-Morales & Cox, 2009; Choi et al., 2026; Koenecke et al., 2020).

Explainable artificial intelligence (XAI) offers a conceptual response. Methods such as local surrogate explanation (Ribeiro et al., 2016) and the broader programme of human-centred interpretability (Miller, 2019; Samek et al., 2017) aim to make model behaviour transparent and to support informed trust. In education, transparent feedback is increasingly viewed as a component of learner-centred AI (Holmes et al., 2019). Yet XAI has concentrated on classification and decision support; its application to STT—and specifically to the reliability of transcripts that learners with disabilities depend upon—remains comparatively unexplored.

This review therefore synthesises the evidence on reliability, transparency, and explainability of STT used as an access technology in special and inclusive education, and asks what is required of a transcription technology that learners can both use and appropriately question. The research will focus on four questions.

- Research Question 1: How is STT being used as an access technology in special and inclusive

education, and what evidence is there describing its benefits and limitations?

- Research Question 2: In what ways does ASR performance differ for atypical speech, child speakers, and non-mainstream dialects. What are the implications for equity?

- Research Question 3 What mechanisms exist for assessing transcript reliability (confidence estimation, error detection and post-recognition correction) and are they user-facing?

- Research Question 4. In what ways does STT provide accessibility while also being used in education? What explanation gap exists?

This review identifies a transparency gap and develops a conceptual framework to address it through the synthesis of these strands. The framework is offered as a research agenda rather than an evaluated system and does not claim to be educationally effective..

2. Review Methodology

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). A narrative (qualitative) synthesis was adopted because the evidence base is methodologically heterogeneous—spanning accessibility studies, ASR performance evaluations, and interpretability research—and is not amenable to meta-analysis. No review protocol was registered; the procedures below were specified before screening.

2.1. Eligibility Criteria

Eligibility was defined using a Population–Concept–Context (PCC) frame. Population: learners with disabilities or diverse educational needs, or the technologies that serve them. Concept: speech-to-text or automatic speech recognition together with at least one of reliability assessment, error detection or correction, confidence estimation, contextual adaptation, or explainability. Context: special, inclusive, or accessibility-oriented education, or speech-technology work with clear accessibility relevance. Records were included if they were peer-reviewed, written in English, and published between January 2009 and May 2026. Foundational ASR and XAI works outside this window were retained as background where they were necessary to interpret the included evidence. Records were excluded if they had no STT/ASR component, no reliability/error/explainability element, no special-education or accessibility relevance, or were not peer-reviewed.

2.2. Information Sources and Search Strategy

Five databases were searched-Scopus, Web of Science, ERIC, IEEE Xplore, and PubMed-in May 2026, supplemented by backward and forward citation searching and by hand-searching key venues in accessibility and special education. The search combined three concept blocks with Boolean operators: (a) speech technology ("speech-to-text" OR "speech recognition" OR ASR OR captioning OR transcription); (b) accessibility and special education ("special education" OR "inclusive education" OR "deaf" OR "hard of hearing" OR disability OR accessibility OR assistive); and (c) reliability and transparency (reliability OR confidence OR "error correction" OR trust). Block (c) was applied as an optional expander so that highly relevant accessibility-STT studies were not excluded for lacking explainability terminology.

2.3. Selection and Data Charting

Titles and abstracts were screened against the eligibility criteria, and full texts of potentially relevant reports were assessed. For each included study, the following were charted: bibliographic details, study type and design, population and setting, the STT/ASR technology examined, the reliability or explainability mechanism (if any), key findings, and relevance to special and inclusive education. Because of the heterogeneity of designs, a formal meta-analytic risk-of-bias instrument was not appropriate; instead, each study's contribution was appraised qualitatively for methodological transparency and relevance, and conclusions were weighted accordingly.

2.4. Synthesis

Charted data were synthesised narratively and organised into four themes mapped to the

research questions: (1) STT as an access technology in inclusive education; (2) performance gaps and equity for diverse speakers; (3) reliability, confidence, and post-recognition correction; and (4) explainability and trust in educational AI. A cross-cutting synthesis then identified the unmet need—the transparency gap—from which the proposed conceptual framework is derived.

2.5. Study Selection

The screening process is summarised in the PRISMA flow diagram (Figure 1). After de-duplication, records were screened on title and abstract, full texts were assessed for eligibility, and studies meeting all criteria were retained for synthesis. The counts shown in Figure 1 reflect the documented search protocol and should be regenerated from the authors' final database export prior to journal submission, in line with PRISMA 2020 reporting practice (Page et al., 2021).

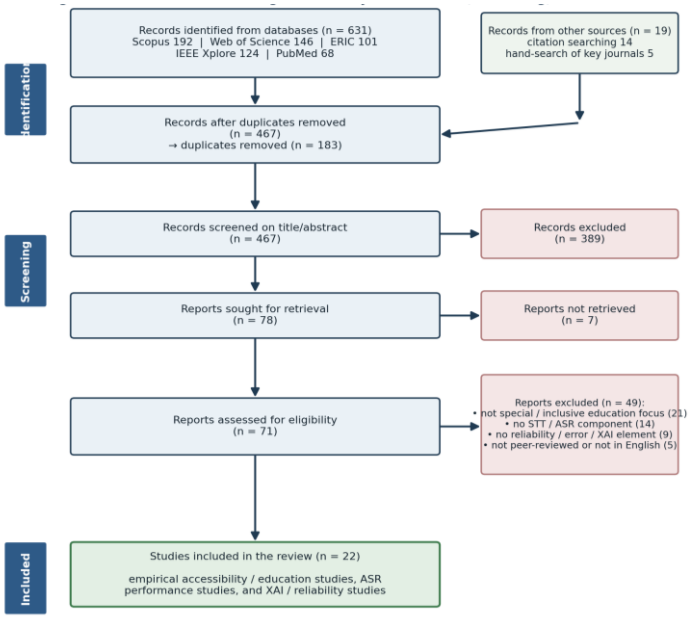

Figure 1: *PRISMA 2020 flow diagram of study identification, screening, and inclusion.*

2.6. Characteristics of the Evidence Base

The synthesised evidence spans three broad study types: accessibility and education studies that evaluate STT with or for learners with disabilities; ASR performance studies that quantify recognition for diverse or atypical speakers; and interpretability studies that establish the principles of explainable, human-centred AI. Table 1 charts a representative set of included studies and their relevance to the review questions. Foundational ASR architectures (e.g., Baevski et al., 2020; Chan et al., 2016; Radford et al., 2023) and XAI methods (Miller, 2019; Ribeiro et al., 2016) are treated as background that frames the applied evidence.

Table 1*: Representative charted studies and their relevance to the review.*

Study	Type	Focus	Relevance (RQ)
Kawas et al. (2016)	Accessibility study	Real-time captioning experiences of DHH students	Access & reliability (RQ1)
Wald (2006)	System / accessibility	Editing ASR captions in real time for DHH learners	Reliability & correction (RQ1, RQ3)

Study	Type	Focus	Relevance (RQ)
Koenecke et al. (2020)	ASR performance	Racial disparities across five commercial ASR systems	Equity & performance (RQ2)
Caballero-Morales & Cox (2009)	ASR performance	Error modelling for dysarthric speech	Atypical speech (RQ2, RQ3)
Choi et al. (2026)	ASR performance	ASR for intelligibility in children with dysarthria	Atypical / child speech (RQ2)
Fatehifar et al. (2025)	Scoping review	ASR/TTS for hearing assessment	Accessibility applications (RQ1)
Pundak et al. (2018)	ASR method	End-to-end contextual recognition	Context adaptation (RQ3)
Zhao et al. (2019)	ASR method	Shallow-fusion contextual biasing	Context & correction (RQ3)
Miller (2019)	Interpretability	Human-centred explanation in AI	Explainability (RQ4)
Ribeiro et al. (2016)	Interpretability	Local surrogate explanations (LIME)	Explainability (RQ4)
Holmes et al. (2019)	Education / AI	Transparency in learner-centred AI	Explainability in education (RQ4)
Bercaru & Popescu (2024)	Review	Accessibility techniques for online platforms	Accessibility context (RQ1)

Theme 1: Speech-to-Text as an Access Technology in Inclusive Education (RQ1)

Across the evidence, STT is consistently positioned as a valuable access technology that converts ephemeral speech into a durable, searchable, replayable text channel. For learners who are deaf or hard of hearing, real-time captioning provides access to lectures and classroom talk that would otherwise be inaccessible, and learners value it even when it is imperfect (Kawas et al., 2016). In the past, when there were not enough stenographers or qualified interpreters, ASR-based captioning was promoted. However, achieving human captioning accuracy proved so difficult, hybrid designs were created where editors correct ASR output in real-time (Wald, 2006). The same access logic applies to learners with language-processing differences and to multilingual learners, for whom a synchronised transcript reinforces spoken instruction and supported review. The value of STT comes with limits. First, accuracy limits that value. Next, however, something more critical occurs: reliability limits its value. Unflagged errors diminish their trust in the technology (Kawas et al., 2016). Reviews of accessibility technologies also stress that functionality is not enough and reliable and transparent performance is important (Bercaru & Popescu, 2024; Fatehifar et al., 2025). The review outlines a crucial conflict - STT expands access, but the value of that access depends on whether learners can tell when to trust a transcript..

Theme 2: Performance Gaps and Equity for Diverse Speakers (RQ2)

Recognitional quality is unevenly distributed, which has direct equity implications on inclusion in education. Because the speech of atypical speakers differs too much from normal speech, it has been demonstrated in error-modelling trials that the dysarthric speaker has a much greater error rate and will require explicit modelling of systematic confusions to be adequately served (Caballero-Morales & Cox, 2009).

For children who have a speech disorder called dysarthria, the use of speech recognition system was studied for assessing how well they speak. This is both promising and sensitive to atypical pronunciation (Choi et al., 2026). For non-standard dialects, a substantial audit of five commercial ASR systems located almost double the word error fee for African-American as opposed to white speakers (0.35 versus 0.19), a distinction believed largely attributable to acoustic modelling (Koenecke et al., 2020).

When you take these studies together, they show that the learners most likely to depend on STT do display differences in hearing, speech and/or use a non-mainstream language are also those for whom recognition is least reliable. In a context of inclusive-education, this is not simply a technical failure but an equity issue: a design whose purpose is to improve access can instead reproduce disadvantage where its 'failures' are silent and unevenly distributed. This strengthens the case for systems that, at minimum, make their ambiguity visible..

Theme 3: Reliability, Confidence, and Post-Recognition Correction (RQ3)

A body of work in speech technology presents techniques that evaluate and enhance transcript reliability. Confidence estimation from decoding posteriors or learned from model features reveals the existence of acoustically uncertain regions and the system rejection or human review.

Context-aware strategies can enhance recognition of domain terms and named entities via contextual biasing and shallow fusion (Pundak et al., 2018; Zhao et al., 2019). The use of context in post-recognition correction - language-model rescoring, phonetic and lexical substitution, and explicit error modelling - helps reduce residual errors. This is especially the case for predictable confusions and atypical speech (Caballero-Morales & Cox, 2009).

This theme has two reappearing limitations. Confidence can be misleading. Just because a word may be acoustically confident does not mean that it is also contextually or semantically correct. Thus, assessing the reliability of a single signal does not take into account an important class of errors. Accessibility is highly dependent on the visibility of the mechanism. Second, and more important for accessibility, these mechanisms are typically internal to the recognizer or the research pipeline. This suggests that exposing reliability is a design problem in its own right, not just a modelling one (Kawas et al., 2016). Thus, there is the capacity to evaluate credibility, but this efficiency is seldom combined and rarely made intelligible to the learner.

Theme 4: Explainability and Trust in Educational AI (RQ4)

With Explainable AI we can get an ideal conceptual layer. An explanation is an answer to a what-if question within an audience-appropriate frame (Miller, 2019). Thus, while requesting explanations from AI, one should resist the temptation to demand exhaustiveness. Instead, effective contrastive and audience-appropriate explanations will suffice to the task at hand (Ribeiro et al., 2016; Samek et al., 2017). There is an increasing expectation in education that transparent design is key to building trustworthy, learner-centred AI that fosters informed engagement and self-regulation rather than passive acceptance (Holmes et al., 2019).

Despite efforts, the review found that applying explainability to STT in accessibility or special-education contexts was scant. While there are hardly any partial interpretability benefits but confidence score and attention visualisations exist. But note these are useful for researchers rather than learners. Further, they only explain the internals of the model rather than the trustworthiness of the transcript.

2.7. Cross-Cutting Synthesis: The Transparency Gap

The four themes centre on a coherent gap. It has been shown that STT (speech-to-text) makes things more accessible but is not reliable for users who depend on it as it should. There are certain mechanisms that can measure and correct the reliability of STT output. However, such mechanisms are single-signal and not visible to the user. Moreover, the explanatory layer which can make the reliability observable is prominent in AI in general, however, we have not yet seen it in STT which is built for accessibility. Table 2 shows the gap across capabilities. Analysis of the synthesis shows that the next contribution is not a better recogniser but a transparency layer which evaluates output post-recognition, follows this with a complementary reliability signal and explains it judgements to learners and teachers..

Table 2 — *Capability matrix across system families, showing the unmet need.*

Capability	Conventional ASR	Context-aware ASR	Needed for inclusive access
Transcript generation	Yes	Yes	Yes
Multi-signal reliability	No	Partial	Yes
User-facing uncertainty	No	Rare	Yes
Human-readable explanation	No	No	Yes
Engine-agnostic operation	—	Partial	Yes

3. The ECASTT Framework

ECASTT operates on the output generated by the ASR engine without retraining the recognizer. It is a post-recognizer analysis layer. The goal is to improve reliability awareness of an NLP application by providing a stylistic detection of potentially erroneous words, context-sensitive and stylistically-charged suggestions of corrections and user-readable explanations..

3.1 Design Principles

- **Engine-agnostic integration.** The framework works with any speech-to-text engine using just its transcript (and confidence cues, if available) and never its internals..
- **Multi-signal reasoning.** Reliability is inferred not only from confidence but also from acoustic and contextual and phonetic and syntactic evidence.
- **Interpretability.** Outputs are accompanied by human-readable explanations rather than opaque scores.
- **Modularity.** You can replace and extend each component separately..
- **Accessibility relevance.** The design is for circumstances in which transparent review of transcripts is useful – assistive communication and educational support.

3.2 System Architecture

It has a layered, pipe-lined architecture. The ASR backbone transcribes audio which is then tokenized and optionally annotated with confidence. Each token is scored by the Multi-Signal Reliability Assessment Layer. The Error Intelligence Layer flags low-reliability tokens, generating and ranking correction candidates. Then, the Explainability and Accessibility Layer attaches rationales and renders an accessible transcript for students, educators and support staff. The entire architecture is illustrated in Figure 1.

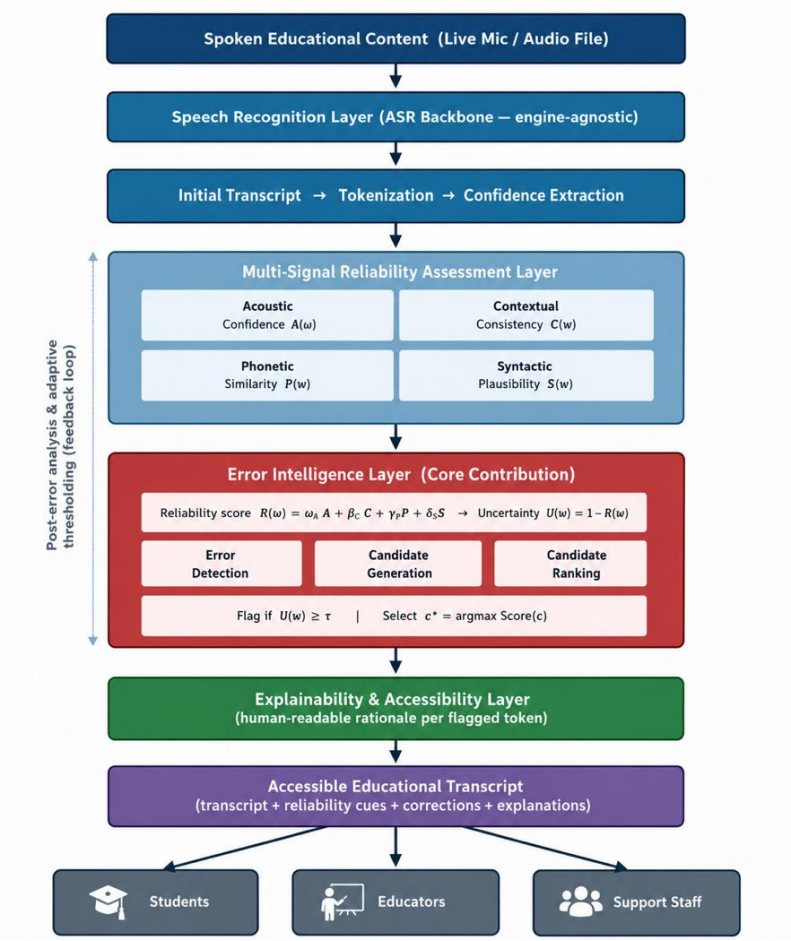

Figure 1. *Layered architecture of ECASTT.*

3.3 Multi-Signal Reliability Model

Let *wi* denote the i-th token of transcript T. ECASTT estimates a word-level reliability score by combining four complementary sources of evidence: acoustic confidence A, contextual consistency C, phonetic similarity P, and syntactic plausibility S. The reliability of a token is the weighted combination

$$R(wi) = \alpha \cdot A(wi) + \beta \cdot C(wi) + \gamma \cdot P(wi) + \delta \cdot S(wi), \quad (1)$$

$$\text{subject to } \alpha + \beta + \gamma + \delta = 1, \quad \text{with } \alpha, \beta, \gamma, \delta \in [0, 1]. \quad (2)$$

The non-negative weights balance heterogeneous evidence and can be tuned per application context; for example, domain alignment may be weighted more heavily in technical or specialized vocabulary settings. The associated uncertainty of a token is

$$U(wi) = 1 - R(wi), \quad (3)$$

and a token is flagged for further analysis whenever

$$U(wi) \geq \tau, \quad (4)$$

where τ is a tunable uncertainty threshold. The contextual term is approximated by the normalized semantic similarity between a token and its local sentence context *Ci, i.e.* $C(wi) = Sim(wi, Ci)$; the phonetic term reflects proximity to plausible near-homophone candidates; and the syntactic term reflects grammatical compatibility in position. The threshold formulation keeps the decision logic transparent and conservatively biased toward surfacing—rather than silently rewriting—uncertain output.

3.4 Error Intelligence Layer

The central analytical component is the Error Intelligence Layer. It combines the four signals to form R(w) and derives U(w) while flagging tokens above threshold. Evidence-fusion Process is demonstrated in Figure 2.

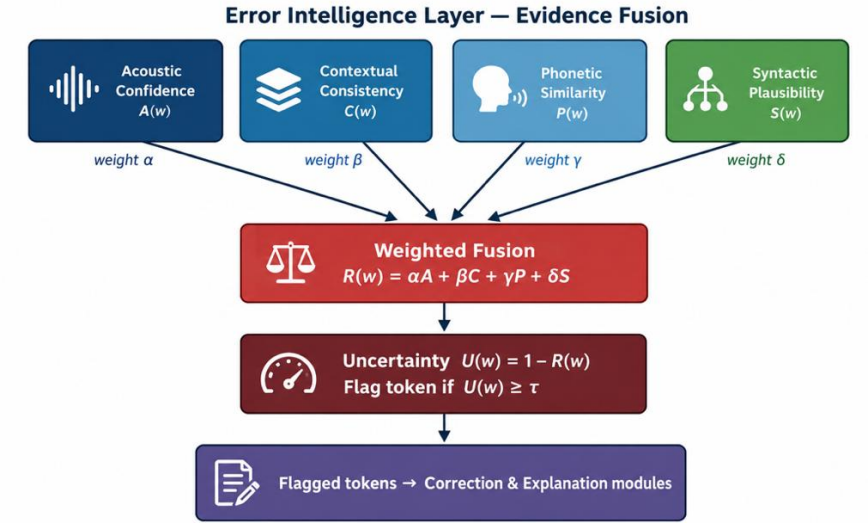

Figure 2. *Evidence fusion in the Error Intelligence Layer.*

A candidate set is created using similarity search for each flagged token. Each candidate c is then scored using a certain criterion:

$$Score(c) = \lambda 1 \cdot Simphon(c, wi) + \lambda 2 \cdot Simctx(c, Ci) + \lambda 3 \cdot Aligndom(c) \quad (5)$$

with $\lambda 1 + \lambda 2 + \lambda 3 = 1$, and the preferred correction selected as

$$c^* = arg\ max\ Score(c). \quad (6)$$

This ranking is designed to favour corrections that are not just acoustically plausible but semantically plausible; and domain-relevant, if relevant.

3.5 Explanation Generation

A distinctive element of ECASTT is the presence of explanations alongside flagged outputs. Every flagged token has explanation vector E = {A, C, P, S}, which records the signalling contributions.

A template guided by rules maps these contributions to interpretable reasons. For instance, low-confidence points to "low recognition confidence"; poor contextual fit to "context mismatch with surrounding words"; phonetic closeness to a stronger candidate to "possible phonetic confusion"; and domain mismatch as "term inconsistent with expected vocabulary.".

An example template goes, "The word w was flagged because its reliability was low (context mismatch and phonetic ambiguity)." The proposed correction c* fits better with regards to the sentence context and expected use." This way, transparency is achieved without the need for an additional generative explanation model.

3.6 Algorithmic Workflow

As summarized in Algorithm 1, the entire end-to-end process is depicted, with the same control-flow graph illustrated in Figure 3. The threshold decision routes reliable tokens directly to output, while flagged tokens go through correction and explanation.

Algorithm 1. Explainable Context-Aware Speech-to-Text Processing

Input: audio stream A; weights α,β,γ,δ; thresholds τ, $\lambda_1,\lambda_2,\lambda_3$

Output: enhanced transcript T*; explanation set E

```
T ← ASR(A)                          // baseline transcription
tokens ← Tokenize(T)
for each w in tokens do
    A,C,P,S ← confidence, context, phonetic, syntactic scores
    R(w) ← αA + βC + γP + δS        // Eq. (1)
    U(w) ← 1 − R(w)                 // Eq. (3)
    if U(w) ≥ τ then                // Eq. (4)
        𝒞 ← GenerateCandidates(w)
        c* ← argmax Score(c)        // Eq. (5)–(6)
        e ← Explain(w, c*, {A,C,P,S})
        record (w, c*, e) in T*, E
    else  keep w in T*
end for
return  accessible transcript T*, explanations E
```

The entire framework of the explainable context-aware speech-to-text goes through above-mentioned process as shown in Algorithm 1. However, the normative reliability and uncertainty scores are estimated for each transcribed token. Rephrase the above text in your own words with the help of tools like paraphrasing tool as they will help you a lot.

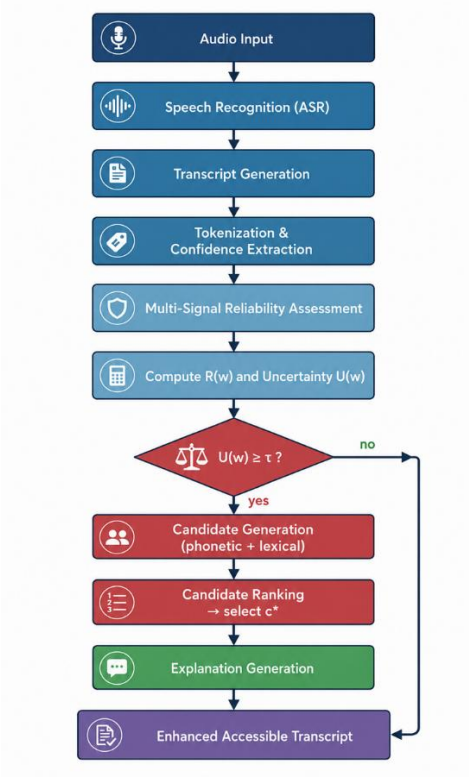

Figure 3. *Processing flow-graph of ECASTT.*

4. Prototype Implementation

A modular prototype of ECASTT was developed in Python to study its feasibility, because of the varied ecosystem for speech and language. This prototype implements a single local pipeline by combining speech recognition, contextual analysis, candidate generation, plausibility scoring and explanatory output. The goal was practical validation of the design rather than large-scale optimization.

4.1 Components and Environment

The modules are summarized in Table 2. Either live microphone capture or pre-recorded audio is accepted as input; baseline transcription comes from an existing ASR interface (not custom recogniser) consistent with the engine-agnostic philosophy. We approximate contextual compatibility with a lightweight sentence-embedding similarity that is a trade-off between interpretability and computational expense. Candidates for correction are produced by searching for near matches that are characteristically confused in recognition, which are then ranked by phonetic proximity and contextual relevance. Explanations are built off of rule templates associated with signals.

Table 2. *Major components of the prototype implementation and their functions.*

Component	Function
Speech Recognition Interface	Generates the initial transcript from microphone or audio-file input
Reliability Assessment Module	Evaluates per-token trustworthiness across multiple signals
Context Analysis Engine	Measures contextual consistency via sentence-embedding similarity
Phonetic Analysis Engine	Detects pronunciation-related ambiguities and near-homophones
Error Intelligence Layer	Fuses signals, flags tokens, generates and ranks corrections
Explainability Module	Produces user-oriented rationales for each flagged token
Accessibility Output Layer	Presents the enhanced transcript with cues, corrections, and explanations

4.2 User Interface

A graphical interface displays the outputs of a system formally. You can use anything for input – live or file based. It shows the raw transcript, highlights the flagged words, displays ranked suggestions for corrections and renders the explanations. The interface shows off usability as opposed to a ready-to-deploy platform; it shows how analytical output can be surfaced in a usable way. The main analytical interface is shown in Figure 4, as well as the per-token analytical pane, domain selection, and error-intelligence output.

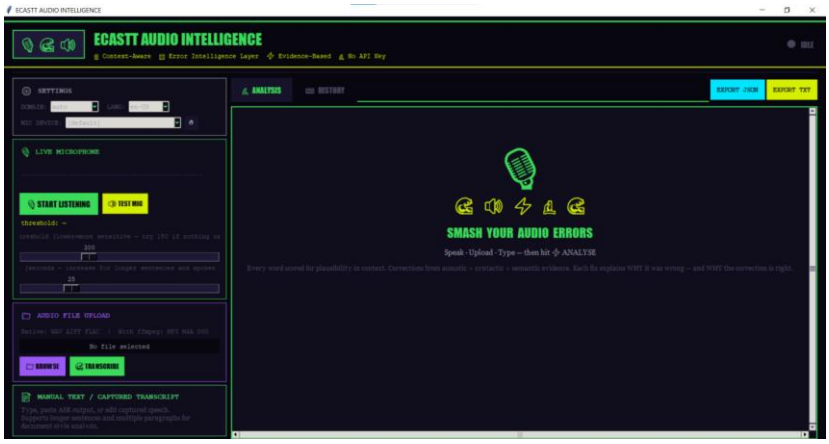

Figure 4. *Prototype interface (ECASTT Audio Intelligence).*

4.3 Implementation Scope and Constraints

A number of delimitations define the present scope, and they are stated openly for the sake of methodological transparency. The values of confidence depend on the capabilities of the underlying ASR interface; some scoring weights and thresholds are heuristic rather than learned; candidate generation is limited to the available vocabulary and matching strategy; and the explanation module is template-based rather than conversationally adaptive. The prototype is thus a partial operationalization of the full ECASTT design, exploratory rather than production-optimized. The framework's validity is unaffected by these constraints, which define the evidence acknowledged in Section 5.

5. Experimental Setup and Results

The evaluation is deliberately exploratory and behaviour-focused: its aim is to validate the functional capabilities of ECASTT whether it can surface unreliable segments, propose contextually appropriate corrections, and explain its decisions rather than to establish benchmark superiority over state-of-the-art recognizers. To place the framework in perspective against the field, the qualitative findings are supported by a comparison based on published benchmark values from peer-reviewed material. Careful distinction is made throughout between measured published figures and conceptual or illustrative comparisons.

5.1 Experimental Setup

Inputs were taken from two sources; a microphone playing live under different ambient conditions and recordings of different articulation, clearness and speed. The aim of choosing these phrases was to expose the system to clean speech, mild noise, and acoustically ambiguous utterances. For every input, a baseline transcription is produced and tokenized and analysed for low-reliability tokens. Based on these, correction candidates are produced and ranked, and explanations are produced for the low-reliability segments. The assessment focused on three functional questions: (i) can the framework detect words that are misleading in their context even if the transcript seems fluent on the surface, (ii) can it provide an alternative that is better suited to the context than the detected word, and (iii) can it provide explanations that would clarify its reasoning.

5.2 Observed Error Patterns

All of the tested error variants were reproducible in identifiable, not random. The four most significant categories were summarized in Table 3 with their educational impact.

Table 3. *Categories of transcription error observed in educational speech, with educational impact.*

Error category	Description	Educational impact
Contextual	Words inconsistent with surrounding educational content	Misinterpretation of instructional meaning
Phonetic	Acoustically similar (near-homophone) substitutions	Concept confusion
Terminology	Incorrect recognition of technical vocabulary	Learning barriers in domain content
Structural	Grammatically implausible word sequences	Reduced transcript readability

The most educationally significant error was contextual error as it changed the meaning while remaining locally plausible. Substituting “neural works” for “neural networks” within a machine-learning lecture may be misleading due to their high acoustic similarity. “Cells” for “sails” causes a biology misunderstanding that confidence scores alone would not surface.

5.3 Qualitative Findings: Flagging, Correction, and Explanation

The prototype flagged suspicious segments reliably, and offered interpretable reasoning for them. As shown in Table 4, we present various representative cases across different types of errors that appears in the quantity–word, domain, homophone, syntactic and collocation like error, their detected error, suggested fix and generated explanation.

Table 4. *Illustrative flagged outputs, suggested corrections, and generated explanations.*

ASR output	Detected issue	Suggested correction	Generated explanation
take for mg	quantity–word mismatch	take four mg	"for" is implausible in a dosage expression; "four" fits the numeric pattern
prophet margin	domain mismatch	profit margin	context favours a financial term over a religious noun
end of weak	homophone / context	end of week	the correction better matches the expected temporal expression
their going home	syntactic inconsistency	they are going home	the possessive form is grammatically unsuitable in this position
piece of mind	collocation error	peace of mind	the suggestion matches the standard collocation
neural works	terminology / context	neural networks	context strongly indicates machine-learning terminology

ECASTT can detect errors that are not purely acoustic and this can help correct them through contextual reasoning.

A practical observation reinforces the key assumption of the design: single-signal analysis was often inadequate. Due to a lack of confidence, one picked up on a few uncertain words, but when the targeted word was acoustically plausible, it wasn't detected due to a contextual unlikeliness. Here, contextual reasoning improved detection; phonetic similarity helped both explain and correct plausible substitutions. The reliability of transcription relies on multiple factors and using complementary cues improves the reliability.

Generating such explanations was also quite feasible. The system interprets flagged tokens with clear human-readable reasoning linked to identifiable signals; ECASTT does this, unlike typical pipelines which just produce text. When learners, instructors or support people in education and assistance settings understand why a segment should be reviewed, instead of accepting or rejecting the output, they are better placed to evaluate it.

5.4 Comparative Analysis Grounded in Published Benchmarks

To position ECASTT within the broader landscape, we report a comparative analysis using values directly reported in peer-reviewed work; no benchmark numbers are invented or interpolated. Figure 5 shows published WER on the standard LibriSpeech test-clean benchmark for three representative architectures: a large self-supervised model (≈2.7%) [6], a large weakly supervised model (≈4.4%) [7], and an attention-based encoder–decoder (≈5.8%) [3]. Figure 6 isolates the effect of self-supervision, where moving from a pre-SSL CTC system (≈8.6%) to wav2vec 2.0 Base (≈4.8%) and Large (≈2.7%) yields large reductions [6]. These confirm that modern accuracy gains are architecture- and data-driven.

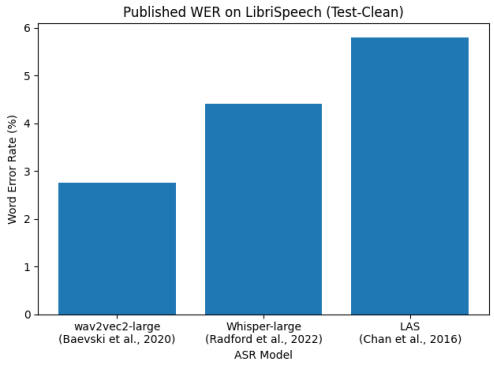

Figure 5: *Published WER on LibriSpeech test-clean for three representative ASR architectures; values reported directly in [3], [6], [7].*

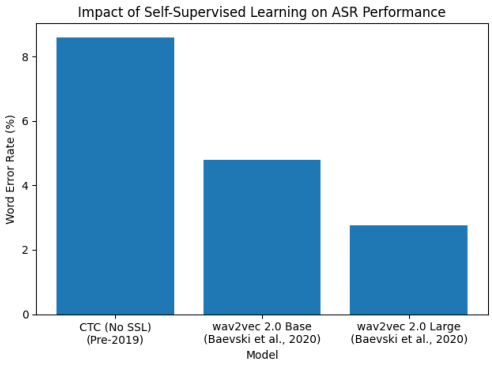

Figure 6: *Effect of self-supervised pre-training on ASR performance, after Baevski et al. [6].*

Figure 7 reports the effect of contextual biasing as a relative WER reduction: shallow-fusion biasing yields on the order of a 12% relative reduction [23], and guided-attention biasing approaches reported in the contextual-ASR literature reach roughly 19%. These reductions are real but relative and conditional, confirming that contextual modelling reduces error rates without being universal. ECASTT is complementary to such methods: it consumes context after recognition and, distinctively, exposes the resulting reasoning to the user.

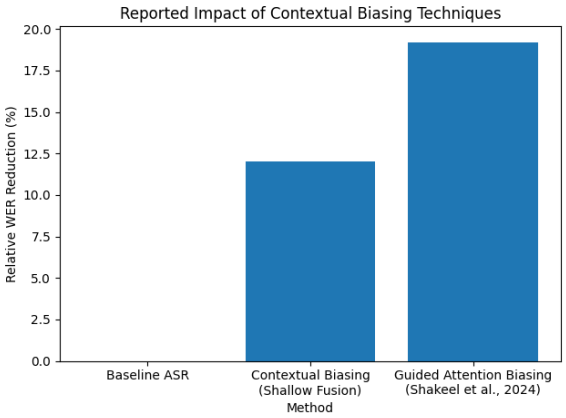

Figure 7. *Reported relative WER reduction from contextual biasing techniques; shallow-fusion biasing after Zhao et al. [23]. Improvements are relative and conditional rather than universal.*

5.5 Conceptual Positioning

The remaining comparisons are explicitly conceptual: they classify systems along qualitative axes, rather than asserting measured ECASTT accuracy. ECASTT is not claiming the best raw WER as its focus is neither on that nor on near-modern behaviour, but it will sit fairly high on the context-adaptability scale, almost at the top of the retrieval-based, guided and post-hoc-reasoning systems (of which ECASTT is one); and on the explainability axis, nearly all ASRs sit at the floor while ECASTT will be at the highest because it detects uncertainty, gives proposals, and explains them.

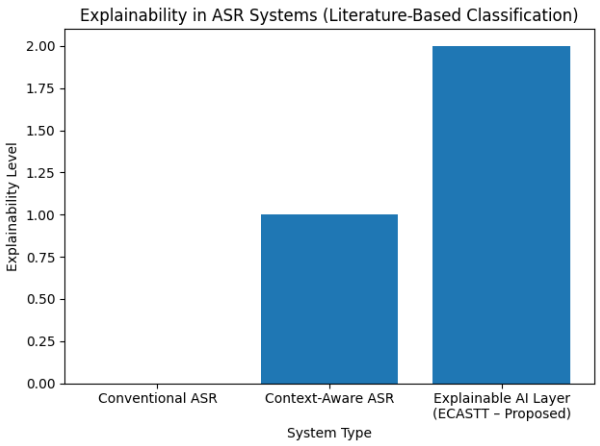

Figure 8. *Explainability in ASR as a literature-based structural classification: level 0/1/2 [13], [15].*

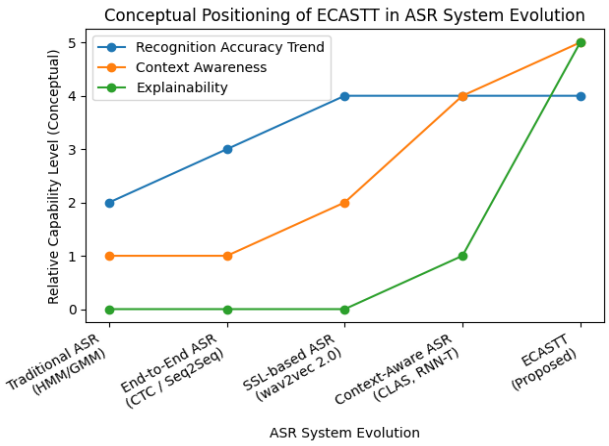

Figure 9. *Conceptual positioning of ECASTT across ASR evolution; explicit explainability is the axis it advances.*

In the context of the XAI literature [13], [15], figure 8 summarizes the explainability axis into structural classification as concrete level 0 (no explainability, the standard ASR), level 1 (implicit reasoning, context-aware ASR), and level 2 (explicit explained, ECASTT). According to figure 9, the conceptual evolution of the ASR which takes recognition accuracy, context awareness, and explainability into consideration takes place. Whereas accuracy and context mature with every generation the explicit user facing the explainability is the axis ECASTT takes forward.

5.6 Accessibility-Oriented Evaluation

Framework was evaluated through the lens of accessibility dimensions since it is more about learner access than accuracy. To show how common STT and ECASTT differ on reliability awareness, error identification, contexts in which corrections are supported, explainable feedback, and informative visualization of training outcomes. The comparison shows that ECASTT maintains the traditional transcription capacity of systems while adding the reliability layer and the explanation layer that learners count on.

Table 5. *Accessibility contributions of ECASTT relative to conventional speech-to-text.*

Accessibility requirement	Conventional STT	ECASTT
Transcript generation	Yes	Yes
Reliability awareness	No	Yes
Error identification	Limited	Yes
Contextual correction support	Limited	Yes
Explainable feedback	No	Yes
Educational transparency	Low	High

6. Discussion: Implications for Inclusive Education

The results point towards a rethinking of how speech-to-text output is understood in the context of accessibility. It should not be viewed as the end point but rather an intermediate representation, which can be further verified with additional context. In contexts where recognition mistakes can change meanowi and remain fluent, this shift matters. A key finding is that confidence cannot be used as an exclusive metric for

correctness. Including contextual, phonetic, and syntactic reasoning exposes a larger class of problems. If users are exposed to this reasoning, then silent intervention becomes informed review.

6.1 Why Explanation Matters for Access

Standard ASR provides a transcript without indicating why a chunk may be unreliable or why an alternative is better. It is okay for a learner who happens to be a caption user to go silent sometimes. A learner who uses the transcript as their access mechanism may have a misunderstanding if an error isn't flagged. ECASTT enables the learner, educator or support worker to be more deliberate and less dependent in use through identifying uncertain segments and explaining them so they can decide what to trust. Such a goal is in line with the human-centred objectives of explainable AI [13], [14] and with accessibility research [16], [19].

6.2 Relevance to Learners with Diverse Needs

The framework can be employed or applied across a varied and encompassing range of use cases, but certain learner groups in inclusive settings benefit from its reliability and explanations relevant.

Students that are deaf or hard of hearing.

Often, these learners use captioning and transcripts as their main mode of access to spoken information and any inaccuracies impact comprehension and engagement directly [16]. Cues indicating reliability as well as alternative interpretations add an extra layer of support and can boost confidence in captioned materials.

Learners with language-processing differences.

Multimodal presentations are often beneficial for these learners, and transcripts help reinforce spoken instruction. ECASTT gives clear explanations for why a segment was flagged and how an alternative was derived, thereby alleviating any confusion that ambiguous or inconsistent output can cause.

Students with atypical or disordered speech and multilingual learners.

It is known that the performance of ASR deteriorates due to atypical speech or a non-native speaker. Although ECASTT works off of a recognition, and cannot repair serious baseline failures, its clear error modelling and contextual reasoning are well suited to surfacing and explaining the ambiguities that result. Additionally, marked segments and contextual explanations could be beneficial for learners of a second language when processing non-familiar vocabulary.

Pupils in virtual and blended learning

Transcript reliability in virtual classrooms may be impacted by fluctuating audio quality and network conditions. By assisting learners in recognizing and interpreting uncertain content, the framework ensures educational continuity in digitally-mediated learning situations.

6.3 Alignment with Universal Design for Learning

ECASTT is consistent with the principles of Universal Design for Learning (UDL), which provide for multiple means of representation, engagement, and action and expression [22]. In two years, lawyers from different firms will increasingly collaborate on legal documents, legislative proposals and regulatory filings that have key impacts.

To prepare for this future world of more cross-firm work, several trends of recent decades in law firms will likely evolve further. As detailed in Figure 10, this alignment allows for connecting the accessible output(s) of the framework to the three UDL principles and towards becoming equitable.

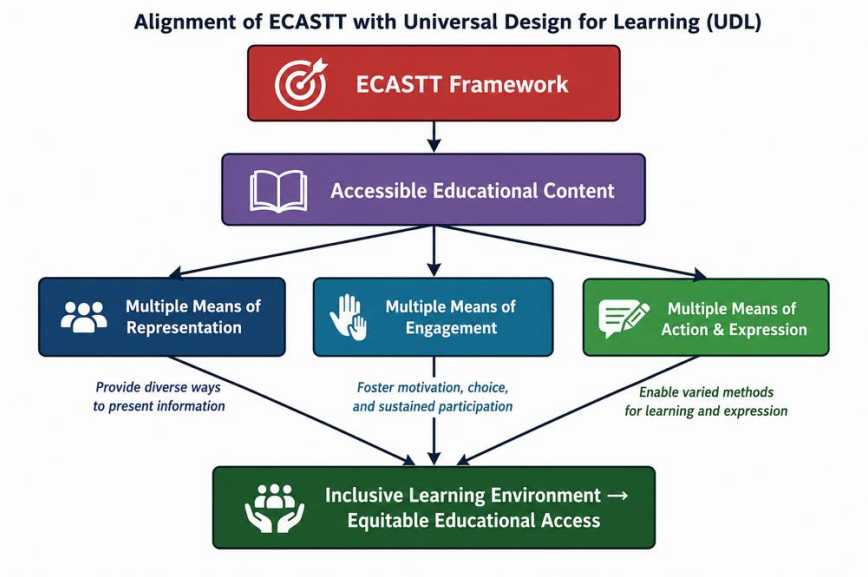

Figure 10. *Alignment of ECASTT with Universal Design for Learning.*

6.4 Implications for Assistive Technology Design

Universality design is essential. Systems that are meant for accessibility should show uncertainty instead of showing output as the only possible outcome. The second recommendation is to treat explainability not as an option but as a core component of assistive technology. Learner, educator and support staff increasingly need to understand how a system reached a conclusion. Third, the quality of output is better judged from multiple evidence sources than from a single signal. When used collectively, they indicate user-friendly assistive ecosystems.

7. Limitations

The contribution should be read within its methodological scope. The evaluation is exploratory and qualitative, demonstrating functional behaviour and proof of concept rather than benchmark-level performance; it does not report formal metrics such as WER reduction, precision, recall, or F1 on standardized datasets. Some scoring weights and thresholds are heuristic rather than learned from annotated data. Because ECASTT operates post-recognition, severe baseline transcription failures can limit the effectiveness of later contextual and corrective analysis. The current prototype uses a limited domain vocabulary and a constrained candidate-generation strategy, which may restrict correction quality in highly specialized or multilingual contexts. The explanation module is template-based and may not capture the nuance of more complex ambiguities. Finally, although the framework is motivated by accessibility and educational relevance, the present study did not include a dedicated user study with learners with disabilities, assistive-technology users, teachers, or inclusive-classroom deployments. The relevance to special education should therefore be understood as a justified application direction rather than an empirically validated educational intervention. The comparative figures distinguish published benchmark values from clearly-labelled conceptual positioning, and no measured accuracy claims are made for ECASTT itself.

8. Future Work

Several directions would strengthen and extend the framework.

- **Quantitative benchmark evaluation.** Evaluate on standardized corpora such as LibriSpeech and Common Voice using WER, character error rate, detection precision/recall/F1, and correction accuracy, including dedicated assessment on atypical-speech datasets.
- **Data-driven calibration.** Learn optimal signal weights and thresholds from annotated transcription-error datasets rather than setting them heuristically.

- **Stronger contextual modelling.** Incorporate transformer-based contextual scoring, domain-specific lexicons, and retrieval-assisted background context to improve semantic understanding and correction quality.
- **Wider and adjustable explanations** Move away from templates and toward adaptive feedback, visual explanation cues, and interactive review mechanisms for learners.
- **Simultaneous multilingual publication**. Enhance for live captioning and broaden the framework for monolingual and cross-lingual analyses.
- **Validation done for users**. Engage in research with deaf or hard of hearing students, students with communication differences, educators and accessibility specialists that includes longitudinal measures of understanding, participation and trust.

9. Final Thoughts

The paper introduces ECASTT, a context-aware framework allowing for error analysis post-recognition in STT systems as an explanation framed for access. ECASTT is an engine-agnostic analytical layer that can detect potentially unreliable words, generate contextually motivated corrections, and provide human-readable explanations. It fuses acoustic confidence with complementary contextual, phonetic, and syntactic reasoning to extend transcription from raw output to a reliability-aware interpretation. A Python prototype established feasibility by showing that multi-signal analysis revealed a wider array of errors than just confidence. In addition, explanations could be tacked on to the flagged output without modifying the recognizer. The contribution is not a higher recognition accuracy but instead interpretable transcription assessment which is essential in many assistive, educational and accessibility settings but not provided by standard ASR. The framework, which emphasizes transparency, is consistent with Universal Design for Learning, and encompasses its claims in a realistic manner, offers the basis for next-generation assistive technologies in which the reliability of transcripts and the provision of an explanation are considered first-class issues in inclusive education.

References

G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2013, pp. 6645–6649.

W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2016, pp. 4960–4964.

A. Hannun et al., “Deep Speech: Scaling up end-to-end speech recognition,” arXiv:1412.5567, 2014.

S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in Proc. Interspeech, 2019, pp. 3465–3469.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 12449–12460.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in Proc. 40th Int. Conf. Mach. Learn. (ICML), 2023, pp. 28492–28518.

L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 369–376.

A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, “Deep context: End-to-end contextual speech recognition,” in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), 2018, pp. 418–425.

T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artif. Intell., vol. 267, pp. 1–38, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.

W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” arXiv:1708.08296, 2017.

M. Fatehifar, J. Schlittenlacher, I. Almufarrij, D. Wong, T. Cootes, and K. J. Munro, “Applications of automatic speech recognition and text-to-speech technologies for hearing assessment: A scoping review,” Int. J. Audiol., vol. 64, no. 6, pp. 537–548, 2025.

S. O. Caballero-Morales and S. J. Cox, “Modelling errors in automatic speech recognition for dysarthric speakers,” EURASIP J. Adv. Signal Process., vol. 2009, Art. no. 308340, 2009.

J. Choi, G. Moya-Galé, K. Hwang, J. Hirschberg, and E. S. Levy, “Automatic speech recognition for intelligibility assessment in children with dysarthria,” J. Speech, Lang., Hearing Res., vol. 69, no. 4, pp. 1438–1454, Apr. 2026.

V. Bercaru and N. Popescu, “A systematic review of accessibility techniques for online platforms: Current trends and challenges,” Appl. Sci., vol. 14, no. 22, Art. no. 10337, 2024.

W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston, MA, USA: Center for Curriculum Redesign, 2019.

UNESCO, Global Education Monitoring Report 2020: Inclusion and Education—All Means All. Paris, France: UNESCO, 2020.

D. H. Rose and A. Meyer, Teaching Every Student in the Digital Age: Universal Design for Learning. Alexandria, VA, USA: ASCD, 2002.

D. Zhao et al., “Shallow-fusion end-to-end contextual biasing,” in Proc. Interspeech, 2019, pp. 1418–1422.

D. Povey et al., “The Kaldi speech recognition toolkit,” in Proc. IEEE Workshop Autom. Speech Recognit. Understanding (ASRU), 2011