



When AI fails to listen: convergent failure patterns across speech impairments and low-resource speech in ASR

Reyhaneh Sohrabi¹

Received: 2 March 2026 / Accepted: 21 April 2026
© The Author(s) 2026

Abstract

Voice-based artificial intelligence (AI) has become deeply embedded in everyday digital interaction, mediating routine activities such as message composition, reminder setting, service navigation, and the control of domestic environments across both personal and professional contexts. Despite the apparent fluency of these systems, instances of failure reveal important limitations in their underlying architectures. This paper identifies a convergence not previously examined: English people with speech impairments and speakers of Persian, two populations with no shared linguistic or physiological profiles that trigger structurally identical failure patterns in ASR systems, including hallucination and language confusion. The review traces both cases to a shared mechanism: distributional exclusion. In the cases examined here, communication breakdowns cannot be attributed to user input: in both instances, the speakers conveyed their intentions clearly; rather, the failure occurred at the level of the model's interpretative processing.

Keywords Automatic speech recognition (ASR) · Disability inclusion · Low-resource languages · Persian · EU AI Act

1 Introduction

Voice-based artificial intelligence (AI) has become deeply embedded in everyday digital interaction, mediating routine activities such as composing messages, setting reminders, navigating services, and controlling domestic environments across personal and professional contexts (Masina et al. 2020; Caroleo 2025). This expansion is commonly justified by the assumption that speech offers a natural and efficient interface, particularly in hands-free or eyes-free situations. The significance of voice-based interaction, however, extends beyond convenience. For many individuals, particularly those with motor or cognitive impairments and speakers of languages with limited technological representation, voice constitutes a primary means of engaging with digital systems, supporting autonomy, reducing reliance on caregivers, and enabling access to essential services (Masina et al. 2020; Pucci et al. 2024).

Despite the frequent framing of voice-based AI as inherently inclusive, persistent structural inequalities continue to shape access to education, employment, and digital services for people with disabilities and speakers of underrepresented languages (Pucci et al. 2024; Parette and Scherer 2004). People with disabilities constitute a highly heterogeneous population, yet assistive technologies are frequently developed around narrow benchmarks of use and capability. Speakers of low-resource languages face a parallel constraint: training data, evaluation protocols, and architectural priorities were established with little or no reference to their linguistic realities. Whether voice-based AI supports or constrains participation therefore depends on how speech variability is accounted for in design, training, and evaluation.

At the technical core of these systems lies Automatic Speech Recognition (ASR), the component that converts spoken input into textual representations and executable commands, determining whether speech is successfully interpreted or effectively excluded. Both commercial and research ASR systems are optimized for mainstream, non-disordered adult speech, embedding assumptions about articulation, fluency, and acoustic regularity that function as implicit system norms (Young and Mihailidis 2010). Recognition performance predictably deteriorates when speech diverges from these assumptions, with accented,

✉ Reyhaneh Sohrabi
Reyhaneh.sohrabi@uab.cat

¹ Department of Translation and Interpreting and East Asian Studies, Universitat Autònoma de Barcelona, Barcelona, Spain

non-standard, and underrepresented varieties consistently receiving lower accuracy, a pattern indicating structural disparities rather than isolated errors (Azizov 2023; Markl and McNulty 2022; Mujtaba et al. 2024). What has received less attention is how these disparities manifest across populations that share no common linguistic or physiological profile. English speakers with speech impairments and speakers of Persian approach ASR from entirely different directions: one through neurologically impaired motor control, the other through a low-resource, non-Latin-script language with limited representation in dominant training corpora. Yet both trigger the same categories of system failure in models such as Whisper, hallucinated outputs bearing no relation to the acoustic input, semantically distorted transcriptions, and language confusion in which the model substitutes a dominant language for the one spoken (Pokel et al. 2025; Iranmanesh et al. 2026). Those two fundamentally different conditions produce structurally parallel failures that expose something neither literature has examined on its own: these failures are organized not around who the speaker is, but around the speaker's distance from the model's training distribution.

This convergence constitutes the central analytical contribution of the present review. Rather than treating speech impairment failure and low-resource language failure as separate technical problems, the review traces both to a shared mechanism: distributional exclusion. When input falls sufficiently far from the patterns a model has learned to decode, the system does not degrade neutrally; it substitutes familiar linguistic structures from its training data, fabricating plausible but unrelated output, defaulting to dominant-language transcriptions, or replacing communicative content with something different.

Accordingly, this review examines how voice-based AI systems handle non-normative speech and linguistically underrepresented languages, with English speech impairments and Persian as its primary cases, by tracing the structural conditions under which recognition failure emerges and identifying the convergent patterns that link these apparently unrelated populations. Empirical evidence on ASR system behavior, including hallucination, semantic distortion, and language confusion, is connected to the design, training, and evaluation choices that produce them, and analyzed not as incidental limitations but as outcomes of how variability is operationalized within current ASR architectures.

Analytically, the review draws on the social model of disability, which locates exclusion in environmental design and institutional practices rather than in individual impairment (Oliver 1996; Public Health Service Ombudsman 2014). From this perspective, barriers arise not from speech variability itself but from systems constructed around narrow norms of intelligibility, reinforcing structural disadvantage by restricting access to the services voice-based AI is designed to mediate (Hawley et al. 2013). Terms such as

non-normative or *atypical* are used throughout this paper to denote divergence from dominant system assumptions, not individual deficit (Public Health Service Ombudsman 2014; Oliver 1990).

The review makes three contributions. First, it identifies a convergence not previously examined: two unrelated populations trigger structurally identical failure patterns traceable to distributional exclusion. Second, it proposes a diagnostic evaluation chain that renders these distinct failure dimensions visible where default metrics cannot. Third, it demonstrates that the regulatory framework closest to addressing these harms, the EU AI Act, lacks the measurement infrastructure to detect them. This review does not conduct original experiments; its contribution is analytical, identifying a convergence across two literatures not previously examined together and proposing the evaluation infrastructure required to test it. The following section situates these failure patterns within the regulatory landscape governing their deployment, examining what the EU AI Act requires, what it omits, and why that gap matters for the populations this review examines.

1.1 AI voice recognition under the EU AI Act

The failure patterns identified in this review, hallucination, semantic distortion, language confusion, produce discriminatory outcomes: a speaker's communicative intent is replaced by fabricated content with no interface signal that anything went wrong. No existing regulatory framework can detect these failures or distinguish them from ordinary transcription errors.

The European Union Artificial Intelligence Act (2024) comes closest. It introduces a risk-based framework that acknowledges discrimination arising from biased or unrepresentative training data, places voice-based interfaces mediating access to employment, education, and essential services under heightened scrutiny, and, through Article 10, requires that training datasets be sufficiently representative and free from biases that could lead to discriminatory outcomes (European Union, 2024).

The gap between this provision and the failures documented here is structural, not only a matter of enforcement. Article 10 locates the source of discriminatory outcomes in dataset composition, framing bias as a problem of representativeness. The convergence identified in this review complicates that framing. English people with speech impairments and Persian speakers show the same three failure types through entirely different causal paths: acoustic atypicality in one case, systemic underrepresentation of a language in the other. What unites them is not a shared data gap but a shared model behavior: when input falls outside the learned distribution, the system defaults to fluent, high-probability output rather than signaling uncertainty. This

distributional default is a property of inference, not of the corpus alone. Article 10(3) does allow for such populations in principle, requiring datasets to reflect the statistical profile of the groups a system is intended to serve. In practice, the clause stays empty: it sets no threshold for when a group counts as covered, no metric for measuring that coverage, and no test for separating a group-specific failure from random error (European Union, 2024). A dataset that satisfies the representativeness criterion would still permit these failures whenever a speaker's input crosses the distributional boundary for any reason.

Two consequences follow. First, regulatory compliance cannot be discharged at the dataset layer; it requires output-level measurement that distinguishes fabrication from mishearing and language confusion from transcription error. Second, Article 10 nowhere names linguistic underrepresentation, low-resource languages, or speech variability as categories of risk, despite requiring that datasets reflect the conditions under which a system is intended to operate (European Union, 2024). Without explicit categorical anchoring and a diagnostic layer at the output level, the Act addresses a narrower mechanism, data underrepresentation, than the harm it is meant to prevent. The evaluation chain proposed in Sect. "Measuring what models hide" supplies this missing infrastructure.

The following section examines this framing through existing empirical evidence, showing how two unrelated speaker populations encounter structurally identical failures when their speech falls outside the model's learned distribution.

2 Different voices, one blind spot

If no two human voices are acoustically identical, each shaped by neurological condition, language background, articulatory anatomy, and communicative context (Simonyan et al. 2016; Garellek 2019), then how do two entirely unrelated speaker populations produce the same failure patterns in the same model?

The choice of these two cases is methodological. English non-normative speakers and Persian speakers offer structurally comparable failures from entirely different origins, one neurological, one institutional. Persian functions here as a proxy for low-resource languages more broadly: the failures it triggers are produced not by properties specific to Persian but by insufficient representation in training data, a condition shared by any language at the periphery of dominant corpora.

Dysarthria¹ presents a severe case. Neurological impairment disrupts motor control over speech musculature, producing instability in articulation, irregular timing, and altered phonation that vary not only across individuals but within the same speaker over time (Young and Mihailidis 2010; Jaddoh et al. 2023). For ASR systems, these characteristics are particularly destabilizing: prolonged or unpredictable pauses trigger hallucinated output where the model fills silence with fabricated text; sound repetitions cause Whisper to hallucinate over 20% of the time (Sridhar and Wu 2025); reduced articulatory precision breaks the temporal alignment on which decoding depends (Caballero-Morales and Cox 2009; Mujtaba et al. 2024).

Persian exposes the same failure from an entirely different starting point. As a non-Latin-script language with limited representation in large-scale speech corpora, it sits at the periphery of Whisper's learned distribution. The lack of large colloquial Persian speech datasets with diverse utterances results in poor automatic transcription; YouTube, for instance, lacks auto-transcription for Persian entirely, leaving a significant gap in real-world evaluation coverage (Iranmanesh et al. 2026). When Whisper encounters Persian input, it produces the same categories of failure documented for dysarthric speech: hallucinated outputs unlinked to the actual audio, semantically distorted transcriptions, and language confusion in which the model defaults to a dominant language rather than the one spoken (Koenecke et al. 2024; Atwany et al. 2025).

The critical distinction is this: dysarthria involves impaired speech production; the acoustic signal genuinely deviates from normative articulatory patterns. Persian speech in these datasets is normative. The speakers are fluent, the articulation is standard, the signal is clean. Yet the model fails on both in structurally identical ways. This means the failure patterns cannot be attributed to signal quality or speaker deficiency.

The structural origin of this convergence lies in what these models are trained on. Dominant ASR corpora such as LibriSpeech² concentrate on fluent, standardized production from younger speakers of American and British English (Mujtaba et al. 2024; Azizov 2023), while commercial systems train on tens of thousands of hours of naturalistic interaction; Google's voice search corpus alone comprises 20,000 h drawn from everyday use (Doshi et al. 2021; Turrisi et al. 2021). Across both benchmark and deployment

¹ Dysarthria is a motor speech disorder caused by neurological impairment, characterized by reduced intelligibility due to disruptions in articulation, phonation, and prosody (Caballero-Morales and Cox 2009).

² LibriSpeech comprises roughly 1000 h of read English speech sampled at 16 kHz, sourced from public-domain audiobooks and segmented for ASR benchmarking (Panayotov et al. 2015).

pipelines, these collection strategies privilege scale, uniformity, and annotation ease over demographic and acoustic coverage, embedding dominant usage patterns as the operational default. Current commercial speech-to-text systems are designed specifically for a mainstream, normative speech population, thus excluding individuals with non-normative speech from the representational foundation on which recognition depends (Young and Mihailidis 2010).

Dysarthric speech datasets remain small, clinically recorded, and limited in utterance diversity (Jaddoh et al. 2023; Turrisi et al. 2021). The consequences for recognition accuracy are severe. While Word Error Rate (WER) for speakers with normative speech may fall below 10%, it can go beyond 50% for speakers with non-normative speech (Tomanek et al. 2024). The disparity holds across systems: one ASR system achieves an average WER of 6.55% on normative speech but 61.90% on dysarthric speech, an absolute difference of 55.35 percentage points on the same recognition task (Turrisi et al. 2021). Whisper demonstrates superior performance among current models but still exhibits a significant drop when disfluencies are introduced, with WER rising from 3.1% to 9.8% (Mujtaba et al. 2024). A consistent and statistically significant accuracy bias exists across all tested ASR systems against disfluent speech, manifesting not only in elevated word-level error rates but also in significant syntactic and semantic inaccuracies that distort the communicative content of the utterance (Mujtaba et al. 2024).

Persian faces a parallel data deficit, though its cause is institutional rather than physiological. Persian accounts for only 392 h in the translation task and 24 h in the transcription task (Namdarzadeh and Ballier 2024). WER reaches 44.8 for Persian versus 6.3 for English on the same model size (Namdarzadeh and Ballier 2024); the original Whisper-small achieves a WER of 91.55 on Persian content without fine-tuning (Pour et al. 2024), and results for Whisper Small are excluded entirely in some evaluations because WER exceeds on noisy entertainment content. Performance scales predictably with model size: the tiny model produces a WER of 2.66 and a CER of 1.22, while the large model achieves a WER of 0.42 and a CER of 0.20 respectively. Comprehensive benchmarks remain scarce despite Persian being spoken by over 100 million people (Iranmanesh et al. 2026). Persian also poses distinct ASR challenges stemming from its linguistic diversity, spanning regional varieties such as Balochi and Kurdish, and features such as variable word boundaries that standard tokenization cannot reliably capture.

Bias introduced at the data level is not merely inherited by the model; it is structurally transformed into architectural behavior. ASR systems learn statistical regularities from their training input, and when those regularities are dominated by normative speech, the learned representations encode normative production as the expected signal. This limitation recurs across model generations, in conventional

hybrid architectures, in end-to-end systems, and in large transformer-based models, indicating that the problem lies in shared assumptions embedded across dominant ASR paradigms rather than in any specific implementation (Doshi et al. 2021; Pokel et al. 2025). Whether explicitly engineered or implicitly learned, core ASR mechanisms presuppose relatively stable speech rates, predictable acoustic boundaries, and consistent realizations across utterances. When this alignment is disrupted by irregular timing, atypical pausing, or inconsistent production, cascading failures follow (Young and Mihailidis 2010; Caballero-Morales and Cox 2009). Even state-of-the-art systems such as Whisper and wav2vec degrade markedly when confronted with non-normative articulation or prosodic variation, a pattern that persists despite increases in model scale (Pokel et al. 2025). Normative speaker adaptation techniques, calibrated for modest variation around normative baselines, offer limited relief precisely where it is most needed, at low levels of intelligibility (Caballero-Morales and Cox 2009). Representational advances alone do not resolve architectural sensitivity to distributional norms.

Critically, the model does not degrade gracefully toward partial recognition under these conditions. It generates two distinct failure patterns: hallucination (2.1), where the model fabricates content unlinked to the actual audio, and language confusion (2.2), where that fabrication takes the form of output in a non-target language.

2.1 Hallucination

Hallucination produces outputs that are sometimes grammatically plausible but semantically disconnected from what was said. Approximately 1% of Whisper transcriptions contain entirely fabricated phrases or sentences bearing no correspondence to the actual audio, and these hallucinations are often non-deterministic, yielding different random text on each run of the API (Koencke et al. 2024). The mechanism is distributional: ASR hallucinations arise when models prioritize distribution patterns over fidelity to audio input, fabricating text unlinked to reference content (Atwany et al. 2025). A speaker who stutters produced the phrase "DRINK FRESH WATER," yet Whisper v2 transcribed it as "Available now", a grammatically coherent but semantically unrelated output that mirrors the formulaic language of online advertisements, suggesting the model defaulted to high-frequency patterns in its training data rather than decoding the actual utterance (Sridhar and Wu 2025). McGuire and Larson-Hall (2025) describe the same tendency more broadly: Whisper selects probable words and reduces grammatical infelicities with no reference to the actual stimulus, optimizing for output plausibility rather than input fidelity (2025). Hallucinations correlate directly with training data composition: phrases strongly associated

with dominant training contexts, such as "Thank you for watching," appear in hallucinated outputs regardless of what was spoken (Atwany et al. 2025). The consequences are not trivial; 38% of hallucinated outputs contain explicit harms, including fabricated associations, misattributed statements, and content that perpetuates violence (Koencke et al. 2024).

The distributional source of hallucination becomes visible when specific populations are examined. For speakers with aphasia,³ extended non-vocal durations, a core feature of the condition, cause the model to treat silence as input requiring completion: 1.7% of aphasia audio segments yield hallucinations versus only 0.2% of control segments (Koencke et al. 2024). Longer pauses in spoken speech more generally could result in more hallucinations due to Whisper being seeded by noise rather than speech. For stuttered speech, the disparity is sharper: Whisper v2 hallucinates on 14% of stuttered clips compared to 3.3% of fluent clips, with sound-repetition clips specifically reaching 22.6%, a rate that Whisper v3 reduces to 6.9% but does not eliminate (Sridhar and Wu 2025). Hallucinations unique to stuttered input take particularly harmful forms: Sridhar and Wu further document degrading onomatopoeic outputs, fabricated phrases such as "oink oink oink" and "bah bah bah", that bear no relation to the speaker's intended content and carry an inherent dimension of dehumanization (2025).

For Persian, hallucination follows the same distributional logic through a different entry point. In smaller model sizes, hallucination intensifies: the tiny model exhibits higher frequency of fabricated output, often repeating clusters of characters or sentences (Namdarzadeh and Ballier 2024). Specifically, in smaller models such as the small model, numbers read by the speaker at the beginning of each sentence are often hallucinated into unrelated content (Ballier et al. 2024). Even in the largest models, hallucination persists in attenuated form, with a verb repeated several times at the end of a sentence, suggesting that scale reduces but does not resolve the underlying distributional mismatch (Namdarzadeh and Ballier 2024). The Faster-Whisper model addresses many of these issues and exhibits significantly fewer hallucinations (Sedghiyeh et al. 2025), but the underlying pattern remains: when the training distribution does not represent a language adequately, the model fabricates from what it knows.

2.2 Language confusion

Language confusion operates through the same mechanism, but instead of fabricating content within the target language,

the model substitutes a partially or entirely different language. For non-normative speech in English, Whisper generates non-English transcriptions even when the target language is explicitly specified as English; in these cases, the model is not misrecognizing the language but generating a different one altogether (Koencke et al. 2024). Pokel et al.'s example illustrates this directly: the model transcribed English non-normative speech containing the Japanese place name "Higashirinkan" as the German sentence "Ein Gassi rennt da," crossing two language boundaries in a single fabrication, pattern-matching to the closest structure available in its learned distribution (2025). For Persian, language confusion takes a distinct but structurally parallel form. Overfitting with Arabic data, explained by the partly shared alphabet between the two languages, causes the model to default to Arabic script and lexical patterns when Persian input is ambiguous (Ballier et al. 2024). Multilingual models such as Whisper and SeamlessM4T⁴ occasionally produce transcriptions in the wrong language entirely due to failures in their language identification modules; for instance, tiny Whisper models generate Chinese characters extensively in Persian transcription (Sedghiyeh et al. 2025; Namdarzadeh and Ballier 2024). For a speaker combining speech impairment with a low-resource language variety, these defaults compound: two independent grounds for distributional mismatch operate simultaneously, with no architectural mechanism to identify or mitigate either.

The convergence between these two populations is not merely thematic; it is structurally specific. Both English non-normative speech and normative Persian input trigger fabricated output derived from the model's training distribution rather than decoded from the actual audio. Both conditions trigger language boundary crossing: Whisper produces German output from English dysarthric input and Arabic, Chinese, and Japanese tokens from Persian input, defaulting to better-represented languages in its training data (Koencke et al. 2024; Ballier et al. 2024). And in both cases, increasing model scale attenuates but does not eliminate these patterns. Whisper v3 reduces stuttering hallucination rates, while Persian hallucination shifts from full character clusters in smaller models to verb repetition in larger ones, but persists across all sizes (Sridhar and Wu 2025; Namdarzadeh and Ballier 2024). These are the same distributional mechanisms generating the same categories of failure across conditions that share nothing except their distance from what the model was trained to decode.

³ Aphasia is a language disorder wherein individuals have lowered ability to express themselves using speech and voice (Koencke et al. 2024).

⁴ The Seamless-M4T models, available in two variants, is designed to support multilingual transcription, translation, and speech-to-text processing (Atwany et al. 2025).

3 What failure produces

The failures documented in the preceding section do not remain inside model internals or evaluation logs. They surface wherever affected speakers depend on voice-based systems to communicate, access services, or participate in institutional processes.

The most immediate harm is communicative erasure when a model fabricates content in place of what was spoken. The speaker's intent is not degraded; it is overwritten by material drawn from the model's training distribution, with no interface signal indicating failure (Pokel et al. 2025). This is qualitatively distinct from mishearing. A mishearing is recoverable: the speaker reformulates, the system re-attempts, and both operate within a shared frame of reference, while a fabrication is not. The output reads as plausible transcription, and neither the speaker nor any downstream process can detect that the content was invented rather than decoded (Koencke et al. 2024; Atwany et al. 2025; Tomanek et al. 2024).

These fabrications acquire a different order of severity when transcription carries evidentiary or clinical authority. In medical settings, hallucinated output can introduce fabricated medication lists, alter patient details, or misrepresent family relationships within automated clinical notes, errors that compromise clinical safety and patient privacy, particularly when downstream practitioners treat the transcription as part of the formal record (Koencke et al. 2024). In legal contexts, fabricated phrases in courtroom or prison call transcriptions can yield decisions grounded in statements the speaker never made (Koencke et al. 2024). What makes these fabrications especially dangerous is their surface coherence: they appear to be fluent, well-formed language, indistinguishable from accurate transcription without access to the original audio. Atwany et al. (2025) confirm that hallucination poses a danger in healthcare, legal, and aviation domains, where conventional metrics fail to detect fabricated outputs carrying life-threatening consequences.

The harm can be extended into employment. AI-driven hiring systems use speech-to-text to transcribe video interviews, and the resulting transcriptions are processed by automated tools to assess candidate qualifications (Koencke et al. 2024). When the transcription stage hallucinates, the downstream assessment operates on a record that does not represent what the candidate said. For speakers with speech impairments, this converts a disability-related characteristic into a measurable disadvantage within a system that presents itself as objective (Mujtaba et al. 2024). For Persian speakers in multilingual hiring processes, language confusion can render responses unintelligible to the automated assessor, not because the speaker was unclear, but because the model defaulted to a better-represented language as Koencke et al.

(2024) confirm that 38% of fabricated transcriptions contain harmful or concerning content. These failures simultaneously withhold opportunities from affected speakers and misrepresent them within the systems that claim to serve them—patterns constitutive of systematic disadvantage in which the distributional structure of training data translates into material harm along lines of disability, language, and identity.

Even where individual incidents might be corrected, the cumulative weight of recurring failure reshapes the relationship between speaker and system. Repeated misrecognition generates cycles of repair, reformulations, workarounds, re-attempts, that impose an interactional burden invisible to aggregate performance metrics (Masina et al. 2020; Pucci et al. 2024). Time, fatigue, and increased reliance on caregivers become the price of participation, costs borne by speakers rather than designers (Hawley et al. 2013). This burden has a threshold beyond which users withdraw. Pucci et al. (2024) report that recurring failure erodes trust across tasks, while Parette and Scherer (2004) document that persistent misrecognition contributes to stigma and heightened social visibility, leading users to disengage in public contexts. The speakers most affected, those whose speech diverges most from training assumptions, are therefore also the most likely to abandon the technology.

What emerges is a consistent pattern: when voice-based systems function as gateways to healthcare, legal proceedings, employment, and essential services, misrecognition operates not as a technical limitation but as a barrier to participation, patterned, predictable, and traceable to the distributional mechanisms examined throughout this review.

A system may satisfy formal ethical criteria while systematically excluding the populations it claims to serve. These harms would pass a WER-based evaluation and an EU AI Act compliance audit,⁵ which raises the question the following section addresses: what would evaluation need to become to make these failures visible?

4 Measuring what models hide

Word error rate (WER) and character error rate (CER) remain the default evaluation metrics for automatic speech recognition. Both operate on the same logic: they align a system's output against a reference transcript and count

⁵ Under the EU AI Act, high-risk AI systems undergo conformity assessments covering training data governance, risk management, and non-discrimination (European Union, 2024). These assessments prescribe no diagnostic tools for transcription quality. WER, the default metric, captures surface error but cannot distinguish mishearing from fabrication, a distinction on which the severity of harm entirely depends.

surface-level mismatches. Substitutions, deletions, and insertions are expressed as a proportion of the total reference length. For Persian specifically, CER provides diagnostic value that WER cannot: it accommodates the spacing inconsistencies characteristic of Persian script, where variation in letter-spacing affects tokenization, functioning effectively as a subtoken-level error measure (Namdarzadeh and Ballier 2024). For decades, this surface-level comparison has served as the primary indicator of ASR performance across languages, populations, and domains.

The limitation of this approach is not that it measures the wrong thing, but that it measures only one thing. WER and CER assume that errors are local mismatches between predicted and expected tokens, a substituted word, a missed syllable, and an inserted filler. When a model operates within its training distribution, this assumption holds: the model mishears, and the metric captures the distance between what was said and what was transcribed. When the model operates outside its training distribution, however, the nature of failure changes. The model no longer simply mishears; it fabricates content unrelated to the audio input, distorts the meaning of what was said, and produces output in languages the speaker was not using. WER compresses all of these into a single number. A phonetic substitution and a fully hallucinated sentence receive the same treatment: both are counted as token-level mismatches. Atwany et al. (2025) expose this directly, demonstrating that high WER can coexist with low hallucination rates while low WER can conceal dangerous hallucinations, a dissociation that makes surface metrics unreliable as proxies for transcription quality, particularly where it matters most.

Two populations illustrate this problem with clarity. For speakers with dysarthria and other non-normative speech, WER is significantly higher compared to normative speech (Tomanek et al. 2024). For Persian, the large model achieves a WER of 0.32 on clean podcast speech but 0.70 on heterogeneous entertainment content (Iranmanesh et al. 2026). In both cases, the elevated WER and CER signal that the model struggles. What it does not signal is *how* the model fails: whether it mishears, fabricates, loses meaning, or switches languages. Answering that question requires evaluation tools that go beyond surface comparison, and it is precisely this question that determines whether a transcription error is a minor inconvenience or a complete communication breakdown.

4.1 Anatomy of a failure

If WER tells us that a model has erred but not how, the challenge is to build an evaluation chain where each tool reveals what the previous one could not. The metrics examined below are not alternatives to WER or CER; they are successive layers of diagnosis, each exposing a dimension

of failure that surface comparison leaves invisible. Traced through concrete cases from both non-normative English speech and Persian, this chain demonstrates how the same model can produce errors that are quantitatively similar at the surface level yet qualitatively different in their nature, origin, and communicative consequence.

4.1.1 From surface distance to semantic damage

The first gap that WER and CER leave open is whether meaning survived the transcription. A model may produce a high error count while preserving the speaker's intent, or it may produce a low error count while quietly replacing what the speaker meant with something entirely different. BERTScore addresses this gap by computing semantic similarity between the predicted and reference texts using contextualized embeddings. Rather than counting mismatched tokens, these metrics measure whether the transcription occupies the same region of meaning as the original utterance. Mujtaba et al. (2024) apply this logic specifically to disfluent speech, extending ASR evaluation to semantic accuracy on the grounds that meaning preservation, not token overlap, is what determines whether communication succeeds. When Whisper transcribes the Persian noun لطف (/lotf/, meaning grace) as لوتف (/lutf/), a sequence that does not exist in Persian, CER registers a single-character difference. BERTScore registers a deeper failure: the output carries no recoverable meaning, because the transcribed form is not a valid word in the language. The surface metric sees a minor error; the semantic metric sees destroyed content (Namdarzadeh and Ballier 2024).

4.1.2 From semantic damage to fabrication

Knowing that meaning was lost does not reveal why. A semantic mismatch could reflect acoustic mishearing; the model confused two phonetically similar words, or something structurally different: the model fabricated content that was never present in the audio. These are not equivalent failures, and BERTScore alone cannot distinguish them. The Hallucination Error Rate (HER) introduced by Atwany et al. (2025) addresses this gap directly. Defined as the ratio of hallucinated outputs to the total number of evaluated examples, HER measures how often a model produces transcriptions that bear no relationship to the actual audio input, outputs that are internally coherent but entirely disconnected from what was said. This distinction matters because, as Sedghiye et al. (2025) observe, evaluation metrics that go beyond surface-level token comparison offer a meaningfully more accurate picture of transcription quality, particularly in cases where standard metrics register only aggregate error without capturing the nature of the failure. The value of HER becomes concrete in cases where fabrication is

severe: Koenecke et al. (2024) document a transcription of aphasic speech in which Whisper produced a violent narrative, including references to killing, entirely absent from the source audio, while simultaneously inserting the non-sensical character sequence "yK pīH" into what should have been standard English text. A WER score registers this as a block of token mismatches; HER identifies it for what it is, the model constructing output from its training distribution rather than decoding the acoustic signal. WER and CER register high errors; BERTScore registers near-zero semantic similarity, but HER reveals what neither surface nor semantic metrics can: this is not a mishearing and not a minor distortion. In both populations examined in this review, the model is not failing to hear but actively fabricating when input falls outside what it has been trained to recognize.

4.1.3 From fabrication to language boundary violation

Hallucination and language confusion are related but distinct failures, and a complete evaluation must separate them. A hallucinated output may remain within the expected language, producing plausible but fabricated content. A language boundary violation occurs when the model produces output in a language the speaker was not using, a failure that is invisible to both WER, which counts token mismatches regardless of language, and HER, which measures fabrication rates but does not verify whether the output belongs to the target language at all. The Word-level Pass Rate (WPR), developed by Marchisio et al. (2024), fills this gap by checking whether every word in the output belongs to the expected target language. Koenecke et al. (2024) provide a case that exposes the limits of existing metrics on this exact point: when processing English-language audio, Whisper produced the Chinese string "谢谢观看," a phrase meaning "Thanks for watching" that likely originates from YouTube metadata in the training corpus. The transcription is internally grammatical and semantically complete, yet it belongs to a language the speaker never used. WER treats each Chinese character as an individual token mismatch; HER may flag the output as a hallucination but does not identify the script switch itself. WPR, by contrast, would reject every token in the output as falling outside the target language, capturing what the other metrics structurally miss. Ballier et al. (2024) trace the mechanism further through subtoken analysis, identifying Arabic and Japanese subtokens embedded within Persian transcriptions, evidence that the model's internal language boundaries become porous when operating at the edges of its training distribution. Sedghiye et al. (2025) confirm that multilingual models such as Whisper may produce transcriptions in the wrong language entirely due to failures in their language identification modules. WPR detects these violations at the output level, a verification step that no standard ASR evaluation pipeline currently

performs, and one that is essential for any language where sparse training data is sufficient to destabilize the model's language identification.

4.2 What evaluation must become

The diagnostic chain traced above exposes a structural gap in current ASR evaluation practice. Each metric in the chain reveals something its predecessors cannot: BERTScore shows that meaning was destroyed where WER saw only moderate error; HER shows that the destruction was fabrication, not mishearing; and WPR shows that the fabrication crossed a language boundary, confirming that the result was a total communication breakdown. No single metric captures this trajectory. Taken individually, each provides a partial view. Taken together, they compose a complete diagnostic of how ASR systems fail when confronted with speech that falls outside their training distribution.

The metrics themselves are not new. HER (Atwany et al. 2025), BERTScore (Muftaba et al. 2024), and WPR (Marchisio et al. 2024), each exists in the current literature. Atwany et al. (2025) demonstrate the value of combining WER, CER, and HER within a single evaluation protocol, establishing precedent for multi-metric ASR evaluation. What does not yet exist is their integration into a unified framework that extends to semantic, language-boundary, and human-grounded dimensions, motivated by a coherent analytical principle. This paper argues that distributional exclusion provides that principle, because the failures it produces operate simultaneously across surface accuracy, semantic fidelity, output fabrication, and language boundary integrity, no single metric and no single dimension of evaluation can detect them. The framework proposed here is therefore not a new metric but an evaluation logic: the minimum structure required to make distributional exclusion visible across both non-normative English speech and low-resource languages such as Persian.

Operationalizing this chain requires embedding it at specific points in the development cycle. During pre-deployment evaluation, developers would run the full sequence on held-out subsets representing non-normative and low-resource speech, with failure at any stage (semantic collapse on BERTScore, fabrication detected by HER, language violation flagged by WPR) triggering targeted intervention before proceeding. To ensure that observed failure patterns reflect distributional mismatch rather than signal quality differences, any comparative evaluation must control recording conditions across both populations.

This has implications that extend beyond methodology. If regulatory frameworks such as the EU AI Act require fairness auditing and risk assessment for high-risk AI systems, the evaluation infrastructure must be capable of detecting the failure patterns those systems produce. An audit that relies

solely on WER will certify as adequate a system that fabricates plausible sentences, destroys communicative intent, and crosses language boundaries, provided it does so at a rate that keeps the aggregate error score within an acceptable range. The diagnostic chain outlined here does not resolve this regulatory challenge, but it establishes its precondition: distributional exclusion cannot be governed if it cannot be measured, and it cannot be measured with tools designed for a fundamentally different kind of failure.

5 Conclusion

A speaker with dysarthria says, “drink fresh water” and the model transcribes “available now.” A Persian speaker produces clean, fluent speech and the model outputs Chinese characters. Neither speaker failed to communicate. The model failed to listen, and in both cases, it failed in the same way.

This review examined how ASR models produce structurally identical failure patterns, hallucination and language confusion, across two populations with no shared linguistic or physiological profile. The answer lies in distributional exclusion when the input falls sufficiently far from what a model has learned to decode; the system does not degrade proportionally to signal difficulty. It fabricates from its training distribution: generating unrelated content, defaulting to dominant-language transcriptions, replacing communicative intent with distributional artifacts. Dysarthric speech reaches this threshold through neurologically impaired motor control; Persian reaches it through institutional underrepresentation. The triggers differ; the mechanism does not.

What gives this convergence its analytical force is what it rules out. Dysarthric speech genuinely deviates from normative articulatory patterns; model failure could plausibly be attributed to signal degradation. Persian speech in these contexts is fluent, clearly articulated, acoustically clean. Yet the model fails on both in structurally parallel ways. Signal quality and speaker deficiency cannot explain this. Only the training distribution, and the exclusions it encodes, can.

Those exclusions propagate across the pipeline. Data collection constraints keep both populations at the margins of training corpora, while Word Error Rate compresses hallucination, semantic collapse, and language boundary violation into a single aggregate score, rendering the most consequential failures invisible. The diagnostic chain proposed here, integrating WER, CER, BERTScore, HER, and WPR, provides the layered evaluation logic necessary to make each failure dimension individually diagnosable. The EU AI Act requires representative datasets and fairness auditing for high-risk systems but offers no mechanism to detect what unrepresentative data produces. A compliance audit built on aggregate accuracy will certify a system that fabricates sentences, destroys

meaning, and crosses language boundaries, so long as the overall error rate holds. The failure mechanisms documented here give empirical specificity to what the Act frames only in general terms.

Three limitations bound these claims. The convergence is argued analytically from existing literature, not validated experimentally; the parallel failures are drawn from separate studies with different datasets, conditions, and protocols. Until both populations are tested under identical control conditions, the claim remains an inference. The convergence is also asymmetric in its evidentiary base. Non-normative speech failures are documented across multiple peer-reviewed studies with controlled conditions and failure-type disaggregation; Persian failures are documented in a smaller and more recent literature. The claim of structural parity therefore rests on firmer ground on one side than the other and should be revisited as peer-reviewed Persian ASR benchmarks mature. The diagnostic chain, likewise, is a conceptual contribution: interaction effects between metrics, intervention thresholds, and cross-linguistic deployment feasibility remain to be established before it can serve the regulatory function this paper argues it must fulfill.

That a neurological condition and a typologically distinct language produce the same categories of failure, one through a genuinely atypical signal, the other through a perfectly standard one, confirms that ASR exclusion is not a response to difficult input. It is a structural property of what the model was built to recognize. Until that structure is governed accordingly, these systems will continue to determine whose speech counts as input, and whose does not.

Acknowledgements This research is funded through the ALFIE project, which has received funding from the European Union’s Horizon research and innovation programme under Grant Agreement (N 101177912). This study is part of the article-based PhD thesis of Reyhaneh Sohrabi in the DTIEAO at Universitat Autònoma de Barcelona (UAB) within the PhD program in Translation and Intercultural Studies.

Author contributions The author confirms sole responsibility for the following: study conception and design, literature review, data analysis, and manuscript preparation.

Funding Open Access Funding provided by Universitat Autònoma de Barcelona. European Union, 101177912.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Competing interests The authors declare that there are no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes

were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Atwany H, Waheed A, Singh R, Choudhury M, Raj B (2025) Lost in transcription, found in distribution shift: demystifying hallucination in speech foundation models. In: Findings of the Association for Computational Linguistics: ACL 2025, Vienna, pp 23181–23203. <https://doi.org/10.18653/v1/2025.findings-acl.1190>
- Azizov D (2023) Voice, accent, and identity in AI interpreting: toward more inclusive language models. *IRE J Icon Res Eng J* 7(6):498–506
- Ballier N, Burin L, Namdarzadeh B, Ng SB, Wright R, Yunès J-B (2024) Probing Whisper predictions for French, English and Persian transcriptions. In: Proceedings of the 7th international conference on natural language and speech processing (ICNLSP 2024), Trento, pp 129–138
- Caballero-Morales SO, Cox SJ (2009) Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP J Adv Signal Process* 2009:308340. <https://doi.org/10.1155/2009/308340>
- Caroleo L (2025) Do we need a voice methodology? Proposing a voice-centered methodology: a conceptual framework in the age of surveillance capitalism. *Societies* Basel 15(9):241. <https://doi.org/10.3390/soc15090241>
- Doshi R, Chen Y, Jiang L, Zhang X, Biadys F, Ramabhadran B, Chu A, Rosenberg A, Moreno Mengibar PJ (2021) Extending Parrottron: an end-to-end, speech conversion and speech recognition model for atypical speech. In: ICASSP 2021 – IEEE international conference on acoustics, speech and signal processing, Toronto, pp 6988–6992. <https://doi.org/10.1109/ICASSP39728.2021.9414644>
- European Union U (2024), Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series
- Garellek M (2019) The phonetics of voice. In: Katz W, Assmann P (eds) *The Routledge handbook of phonetics*. Routledge, London and New York, pp 75–106
- Hawley MS, Cunningham SP, Green PD, Enderby P, Palmer R, Sehgal S, O'Neill P (2013) A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Trans Neural Syst Rehabil Eng* 21(1):23–31
- Iranmanesh R, Ziaei R, & Garman J (2026) Segmentation strategy matters: benchmarking Whisper on Persian YouTube content. In: Proceedings of the first workshop on NLP and LLMs for the Iranian language family, Rabat, pp 121–130. <https://doi.org/10.18653/v1/2026.silkroadnlp-1.13>
- Jaddoh A, Loizides F, Rana O (2023) Interaction between people with dysarthria and speech recognition systems: a review. *Assist Technol* 35(4):330–338. <https://doi.org/10.1080/10400435.2022.2061085>
- Koenecke A, Choi ASG, Mei KX, Schellmann H, Sloane M (2024) Careless Whisper: Speech-to-text hallucination harms. In: Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (FAccT '24), New York, pp 1672–1681. <https://doi.org/10.1145/3630106.3658996>
- Marchisio K, Ko W-Y, Berard A, Dehaze T, Ruder S (2024) Understanding and mitigating language confusion in LLMs. In: Proceedings of the 2024 conference on empirical methods in natural language processing, Miami, pp 6653–6677. <https://doi.org/10.18653/v1/2024.emnlp-main.380>
- Markl N & McNulty SJ (2022) Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In: Proceedings of the thirteenth language resources and evaluation conference, Marseille, pp 6328–6339
- Masina F, Orso V, Pluchino P, Dainese G, Volpato S, Nelini C, Mapelli D, Spagnolli A, Gamberini L (2020) Investigating the accessibility of voice assistants with impaired users: mixed methods study. *J Med Internet Res* 22(9):e18431. <https://doi.org/10.2196/18431>
- McGuire M, Larson-Hall J (2025) Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: steps toward automation. *Res Methods Appl Linguist* 4:100197. <https://doi.org/10.1016/j.rmal.2025.100197>
- Mujtaba D, Mahapatra N, Arney M, Yaruss J, Gerlach-Houck H, Her-ring C, Bin J (2024) Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In: Proceedings of NAACL 2024: Human language technologies. Association for Computational Linguistics, pp 4795–4809
- Namdarzadeh B, Ballier N (2024) Audio LLM subtokens as encapsulated “knowledge”: The case of Persian subtoken graphemic representations in Whisper. In: Grapholinguistics in the 21st century—from graphemes to knowledge, Venice
- Oliver M (1990) *The politics of disablement*. Macmillan, London
- Oliver M (1996) *Understanding disability: from theory to practice*. Macmillan, London
- Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5206–5210
- Parette P, Scherer M (2004) Assistive technology use and stigma. *Educ Train Dev Disabil* 39(3):217–226
- Pokel N, Moure P, Boehringer R, Liu S-C, Gao Y (2025) Variational low-rank adaptation for personalized impaired speech recognition. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2509.20397>
- Pour MHR, Rastin N, Kermani MM (2024) Persian automatic speech recognition by the use of Whisper model. In: 2024 20th CSI international symposium on artificial intelligence and signal processing (AISP), Babol, pp 1–7. <https://doi.org/10.1109/AISP61396.2024.10475300>
- Pucci E, Piro L, Possaghi I, Mulfari D, Matera M (2024) Co-designing the integration of voice-based conversational AI and web augmentation to amplify web inclusivity. *Sci Rep* 14(1):16162. <https://doi.org/10.1038/s41598-024-66725-3>
- Sedghiyeh N, Sadeghi S, Khodadadi R, Kashani F, Aghdaei O, Rahimi S, Safari MS (2025) PSRB: a comprehensive benchmark for evaluating Persian ASR systems. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2505.21230>
- Simonyan K, Ackermann H, Chang EF, Greenlee JD (2016) New developments in understanding the complexity of human speech production. *J Neurosci* 36(45):11440–11448. <https://doi.org/10.1523/JNEUROSCI.2424-16.2016>
- Sridhar C, Wu S (2025) J-j-j-just stutter: benchmarking Whisper's performance disparities on different stuttering patterns. In: Proceedings of Interspeech 2025, pp 3753–3757. <https://doi.org/10.21437/Interspeech.2025-2700>
- Tomanek K, Tobin J, Venugopalan S, Cave R, Seaver K, Green JR (2024) Large language models as a proxy for human evaluation in assessing the comprehensibility of disordered speech transcription. In: ICASSP 2024 – IEEE international conference on acoustics, speech and signal processing, Seoul, pp 10846–10850. <https://doi.org/10.1109/ICASSP48485.2024.10447177>
- Turrisi R, Braccia A, Emanuele M, Giulietti S, Pugliatti M, Sensi M, Fadiga L, Badino L (2021) EasyCall corpus: A dysarthric speech

dataset. In: Proceedings of Interspeech 2021, pp 41–45. <https://doi.org/10.21437/Interspeech.2021-549>

Young V, Mihailidis A (2010) Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. *Assist Technol* 22(2):99–112. <https://doi.org/10.1080/10400435.2010.483646>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.